MATHEMATICAL MODELS FOR *MYCOBACTERIUM TUBERCULOSIS* COMPLEX GENOTYPING AND PATIENT DATA

By

Inna Vitol

A Thesis Submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
Major Subject: Computer Science

Approved by the
Examining Committee:

Kristin P. Bennett

Dr. Kristin P. Bennett, Thesis Adviser

Dr. Mukkai Krishnamoorthy, Member

Dr. Joyce Diwan, Member

Dr. Mohammed Zaki, Member

Rensselaer Polytechnic Institute
Troy, New York

April 2006
(For Graduation May 2006)

CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Prof. Kristin P. Bennett for her support with this research project. Prof. Bennett has been a wonderful advisor, encouraging, patient, and caring. I would also like to kindly thank Dr. Jeffrey Driscoll of the Wadsworth Center of the New York State Department of Health for providing the genotyping data and very helpful insights on the analysis of the data.

My gratitude is extended to Dr. Natalia Kurepina and Dr. Barry Kreiswirth of the Public Health Research Institute for providing valuable comments. Many thanks to Carla Clark, MPH, and Cynthia Driver, DrPH, of the Bureau of TB Control of the New York City Department of Health and Mental Hygiene for their assistance with the NYC database.

I greatly appreciate the help of Prof. Jeff Trinkle and Prof. Mukkai Krishnamoorthy. Special appreciation to Prof. Michael Danchak is expressed. I am indebted to Prof. Harry W. McLaughlin, Dennis Gornic, and many other wonderful people from RPI. Prof. Sandra Nierzwicki-Bauer, my former advisor, has been very encouraging throughout my years in RPI.

Thank you to all my friends in RPI. Finally, an assistance of the excellent staff of the Computer Science Department, Chris Coonrad, Terry Hayden, and Jacky Carley, is acknowledged.

ABSTRACT

The goal of this project was to develop biologically appropriate mathematical models for genotyping and patient data and use them to analyze and exploit the information in heterogeneous genotyping and epidemiological databases. These databases can be used to address fundamental questions in public health, particularly dynamics of emerging infectious diseases. This work focuses on *Mycobacterium tuberculosis* complex (MTC) because tuberculosis (TB) presents a reemerging serious health threat worldwide; the most optimistic scenarios predict in excess of 80 million new cases and 20 million deaths in the coming decade. Moreover, mycobacteria are one of the most widely sequenced pathogenic groups, and global TB databases currently exist.

Advances in molecular methods contribute significantly to our understanding of the spread of TB. Differentiating between various patent isolates and using the data to guide the efforts of TB control programs are major applications for MTC genotyping. Our research develops mathematical models for spacer oligonucleotide typing (spoligotyping) and demographic data on TB patients. The spoligotyping method exploits polymorphism in the direct repeat locus of chromosome of the MTC bacteria. Spoligotyping produces a simple binary pattern for each TB isolate and is widely used for MTC strain discrimination.

We present SPOTCLUST, a novel mixture modeling approach to advance global studies of MTC genotyping data. SPOTCLUST incorporates biological information on spoligotype evolution without attempting to derive the full phylogeny of MTC. The algorithm is applied to spoligotyping data identified among strains isolated between 1996 and 2004, primarily from New York State TB patients. Our results both confirm previously defined families of MTC strains and suggest certain new families. We demonstrate on New York City demographic data how the resulting models can potentially form the basis of TB control tools using genotyping. Several alternative methods of analysis of MTC genotype and patient data are explored. Improvements to the current method are suggested. Future work will con-

centrate on developing methods for merging probabilistic models for spoligotypes and results from other TB genotyping methods with traditional epidemiological data.

CHAPTER 1
Introduction

This chapter serves to describe the motivation for this work induced by the worldwide tuberculosis (TB) epidemics problem and the increasing role of molecular techniques in the progress toward TB control. The principal TB genotyping methods are presented. The chapter also delineates the approaches developed to date for classification of *Mycobacterium tuberculosis* complex (MTC) strains, and introduces the approach adopted in this work. Finally, the content of the thesis is outlined.

1.1 Motivation

TB is primarily a disease of the respiratory system, but may also affect bones, urinary tract, reproductive and digestive systems, and skin. TB causing bacteria are obligate human and animal pathogens, with a delay between initial infection and the development of clinical disease often up to 5 years. TB is one of the most widespread infectious diseases in the world, infecting more than 2 billion persons, and is expanding due to the HIV/AIDS epidemics and complicated by the emergence of multi-drug resistant MTC strains. One third of the world's population is infected with TB. More than two million people die each year of TB, despite the fact that it is curable with early detection and prompt treatment.

Historically, epidemiologists tracked transmission of TB without knowing the genetic fingerprints of the MTC strains they were following. This greatly increased the difficulty deciphering the spread of MTC strains. If two people have TB strains with matching DNA fingerprints, it is likely that their infections are directly or indirectly linked to each other in a transmission chain. Tracking the source of TB and identifying others who may have been infected becomes much faster and more accurate when strain types are known. Although precise DNA fingerprinting methods have now been developed to identify and track MTC strains, mycobacteriologists in hospitals and public health laboratories do not routinely perform these tests and the bacteria are not genetically characterized. In the mid-1990s, the Centers for Disease

Control and Prevention established a pilot program to evaluate the use of molecular strain typing of MTC isolates in state TB control programs. Molecular typing of MTC strains became an integral part of very successful TB control methods developed by the New York City Department of Health (NYCDOH), the Public Health Research Institute (PHRI), and the Wadsworth Center. Regional and international programs for TB genotyping have been proposed.

Differentiating among various patient isolates and using the data for contact investigations and epidemiological cluster analysis are major applications for MTC strains genotyping. Molecular methods may contribute significantly to classical epidemiological studies, but as MTC genotype databases accumulate data, the tools for analyzing this information do not keep pace. Rapidly expanding national and international databases necessitate development of computational methods to analyze and exploit the large volumes of heterogeneous data. Despite some successful attempts, TB epidemiologists are still in great need for automated analytical and decision-making tools for exploiting genotype databases. Our work presents a first step in an ongoing project dedicated to statistical modeling of MTC genotyping and patient data.

1.2 MTC genotyping

Different genotyping methods are used to generate fingerprints of MTC strains, the most widely employed of which are spoligotyping [66], insertion sequence *6110* restriction fragment length polymorphism (IS*6110*-RFLP) [147], and mycobacterial interspersed repetitive units (MIRU) typing [84, 132, 133]. Another method, also useful for inference of phylogeny of MTC strains, is based on synonymous single nucleotide polymorphisms (sSNPs) [50]. An ideal typing method should be rapid, reproducible, inexpensive, and directly applicable to the problem [149]. A single DNA fingerprinting method sufficiently discriminating for every MTC clinical isolate does not exist [27]. It is highly advantageous to use multiple genetic markers in combination with epidemiological data [83].

The primary goal of genotyping is to discriminate MTC isolates to be able to discern recent transmission of TB versus reactivation of latent infection thus allowing early identification and control of outbreaks. Also, using methods that are reproducible and easily comparable between laboratories allows creation of global databases, which can be extremely useful in tracking transmission routes of TB and can help develop effective protocols for TB control practices. Molecular methods can be efficiently exploited to guide traditional epidemiological approach of a cluster, a collection of MTC strains with identical genotypes, investigation. Cluster investigation is a costly, labor and time-consuming process, which complicates its use in underdeveloped countries with a high prevalence of TB.

This research focuses on developing mathematical models for spoligotyping data. Spoligotyping assay is based on polymorphism in the direct repeat (DR) locus of the MTC bacterial chromosome [66]. The DR locus consists of well-conserved direct repeats interspersed with unique spacer sequences. The region comprising the DR plus the adjacent spacer has been termed the direct variable (or variant) repeat (DVR) [48]. Spoligotyping differentiates MTC strains by determining the absence or presence of 43 defined spacer sequences.

Spoligotyping is a fast and highly reproducible method. The resulting fingerprint has a simple binary format, which permits exchange of data and facilitates construction of large collaborative databases [35, 36]. Octal representation of spoligotypes has also been adopted [22]. Previous studies have grouped spoligotypes into nine major families [113] that can be further broken down into 36 subfamilies in the global database SpolDB3 using visual rules [35]. Groups of related spoligotypes were interchangeably called (sub)families, (sub)clades, and classes [35, 36]. Throughout this thesis, we will most often use the term family for a collection of related spoligotypes. Occasionally, the notation of cluster will be employed. Since this work is of interest for epidemiologists, we specifically indicate when we utilize cluster notation in epidemiological language.

1.3 MTC strain classification

Prior methods for automatic classification of MTC strains into families based on spoligotyping used decision trees induced from the DB1 spoligotype database labeled by a human expert [113]. Decision trees are a form of supervised classifiers, since the families, defined from visual observation of the data, were assumed to be known a priori. "Uninformative" examples were removed using a prototype selection algorithm [113]. While producing interpretable results, the decision tree approach required labeling the data, an error-prone and labor-intensive process compounded by the fact that the phylogeny of MTC families is still under investigation. Moreover, the construction algorithm for decision trees treats spoligotype patterns as simple binary data, not taking into account their biological characteristics, thus oversimplifying MTC strains classification task.

Generative mixture models [100] are a robust form of unsupervised classifiers. Our unsupervised generative mixture models can both identify potential MTC families and create good predictive models for spoligotype classification without requiring labeling and preprocessing. Moreover, this technique can be customized to exploit prior information on TB causing bacteria. Our underlying mixture model assumes that within an MTC family, the spacers can be treated as independent Bernoulli variables, which is the assumption used in the Naïve Bayes classifier. This classifier has been reported to perform surprisingly well despite the deliberately naïve independence assumption [5, 69, 86].

The multivariate Bernoulli mixture model treats features as conditionally independent given the class. Here the features are the absence or presence of spacers. It is widely hypothesized that spoligotypes evolve by deletion of a single or multiple contiguous spacers and that spacer duplication is very unlikely [148, 154]. We have incorporated this knowledge into our algorithm by introducing "Hidden Parents" into the model. We used the expectation-maximization (EM) algorithm to find maximum likelihood (ML) estimates of the mixture model's parameters. Thus, we combined the Naïve Bayes assumption with the EM algorithm, employing one of the promising approaches in unsupervised classification [145].

The performance of the method is greatly dependent on the initialization of the EM algorithm and the number of components in the mixture model. The number of families present in the spoligotype data and the probability distribution for each of them were estimated using Monte Carlo cross-validation (MCCV) technique, which was developed to extract as much information from the data as possible, without any prior knowledge [118]. We used the stability or average best match [55] and log-likelihood to choose a final mixture model. The results were compared to the families identified using prototypes based on the visual recognition rules extracted from the SpolDB3 database [35]. Evaluating the NYC patient data in the context of the identified genotype families proved the usefulness of grouping MTC strains by their spoligotypes while taking into account the suggested direction of their evolution toward the loss of the DVRs.

The results of the application of the Bernoulli mixture models to spoligotyping data confirm some previously defined spoligotyping families [35, 113], as well as identify new families. It may be suggested that some of the prototypes resulting from human-expert-derived visual recognition rules are redundant. Probabilistic methods are well suited for modeling spoligotyping data and should be refined to better fit the current, albeit limited, knowledge on the evolution of the DR genomic locus. Employing the first-order Markov process to model Hidden Parents resulted in an improvement in the quality of the identified families. Future efforts will concentrate on determining patterns of the DVRs' interdependencies and introducing a hierarchy of spoligotyping patterns. The refined models should be enriched with data obtained by using other genetic markers and with epidemiological data; data fusion methods will be developed to efficiently assimilate information contained in these diverse types of data.

1.4 Structure of the thesis

This thesis is organized as follows:

Chapter 2 reviews related work on molecular epidemiology and phylogeny of MTC bacteria as well as prior work on computational analysis of MTC genotyping data.

Chapter 3 provides background on clustering methods, particularly on the mixture model approach. It also introduces the Naïve Bayes assumption underlying our Bernoulli mixture model approach and presents the EM algorithm that is widely used to estimate parameters of mixture models. Cluster quality assurance methodologies are discussed.

In Chapter 4, we describe SPOTCLUST, the proposed approach to cluster spoligotyping data. This chapter includes description of the data used for the analysis, adopted probabilistic framework, newly introduced Hidden Parents that ensure the identification of biologically relevant families, and initialization and validation techniques for our mixture models. We report and discuss the results of the application of our approach to the New York State spoligotyping data.

Analysis of demographic data on New York City patients within the context of the families identified by our method is presented in Chapter 5. Interesting trends were observed while indicating that the probabilistically defined families can provide useful insights into MTC strain data and help guide TB control efforts.

Chapter 6 presents several alternative methods of analysis of spoligotyping and patient data. We showed that joint modeling of patient and genotype data can be advantageous for identification of MTC strain families. Constructing cluster ensembles is suggested as a valuable tool for MTC strain data analysis and visualization. We also demonstrate that graphical models are a natural means to model our data.

Chapter 7 concludes the results presented in this thesis and outlines future directions.

CHAPTER 2
Molecular and computational methods for TB control

In this chapter, we review the genotyping techniques customarily utilized for the objectives of the TB molecular epidemiology. Both advantages and drawbacks of each of the methods are elaborated. Methods of phylogenetic analysis of MTC strains are described. Finally, the current state of the research area concerned with the computational analysis of genotyping data on MTC strains is presented.

2.1 Molecular epidemiology of TB

Ideally, to access genetic variability between bacterial strains, we would sequence and then compare their whole genomes. This, however, is a time-consuming, labor and cost-intensive process, which is impractical for TB control. To overcome these difficulties, only specific genomic loci that bear enough dissimilarity among different strains are used to produce genotype fingerprints of *Mycobacterium tuberculosis* complex (MTC) isolates. These molecular genotyping methods exploit the polymorphism in the number and genomic location of repetitive elements. In general, to be used as a genetic marker, an element should be locus-specific, polymorphic, and easily genotyped.

MTC bacteria are characterized by unusually low polymorphism in structural gene sequences [127]. The fact that there are very few silent nucleotide substitutions in the MTC genome has been interpreted as indicating that MTC is evolutionary relatively young, around 15,000 to 20,000 years old, and that it has disseminated globally recently [127]. MTC has a strongly clonal population structure [4, 127, 135].

Restriction fragment length polymorphism (RFLP) analysis with probes derived from the insertion element IS*6110*, introduced in 1993 [147], is the "gold standard" method for typing MTC strains [93]. IS*6110* is an insertion element that is present in 99% of MTC clinical isolates [67]. All copies of IS*6110* are nearly identical in sequence. However, their copy number varies from 0 to 25 and the location in the genome differs, leading to different RFLP patterns [15]. IS*6110*, once thought to transpose itself randomly within the MTC genome, has now been shown to have preferred sites (hotspots) for integration in the genome [32, 48, 73, 77, 153]. Strains with fewer copies of the IS*6110* have more homogenous fingerprints than do strains with multiple copies of IS*6110* [93]. The frequency of IS*6110* transposition increases with the number of copies of this element [137].

IS*6110*-RFLP is characterized by good discriminatory power and high reproducibility. However, it is labor-intensive, requiring culturing the slow-growing MTC bacteria for several weeks, and is difficult to standardize between laboratories [9]. Moreover, this method does not provide sufficient strain discrimination when fewer than five [126, 149] or too high number [8] of IS*6110*-hybridizing bands are present. Another RFLP-based method is a polymorphic GC-rich sequence-RFLP [17, 101], which has the same disadvantages as the IS*6110*-RFLP [93].

Development of PCR-based genotyping methods has greatly improved typing of MTC strains. PCR-based methods do not require culturing the bacteria, and only small amounts of DNA, which can be obtained directly from the clinical specimen, are sufficient for analysis.

The most widely used PCR-based method is spoligotyping [66]. It is based on the polymorphism in the direct repeat (DR) locus of the mycobacterial chromosome. The DR locus is one of the most well studied loci of the MTC genome showing considerable strain-to-strain polymorphism [32]. The function of the DR locus in MTC bacteria is presently unknown [148]. The well-conserved 36-bp direct repeats are interspersed with unique spacer sequences varying from 35 to 41 bp in size. The order of the spacers was found to be well conserved [148]. The region comprising the DR plus the adjacent spacer has been termed the direct variable repeat (DVR) [48]. Currently 94 different spacer sequences have been identified, of which 43 are used for MTC strain differentiation [148]. Clinical isolates of TB causing bacteria (*M. tuberculosis sensu stricto*, *M. bovis*, *M. africanum*, *M. microti*, and *M. canettii*) can be differentiated by the presence or absence of one or more spacers. Spoligotypes are believed to evolve by deletion of one discrete or multiple contiguous spacers. Various genetic mechanisms, such as homologous recombination, transposition, DNA replication slippage, or point mutation, can cause the deletion [93, 154]. The DR region is one of the hotspots for the IS*6110* integration [48, 77, 148]. The frequencies of deletions in the DR locus were suggested to depend on strain family [154]. Beijing strain family, for example, is extremely stable and appears to be in an evolutionary fixed state [148]. The rate of the evolution of DR variants is slower than that of IS*6110* patterns [154].

Another PCR-based method, increasingly used in molecular epidemiology of TB, employs variable numbers of tandem repeats (VNTR) [42] of genetic elements called mycobacterial interspersed repetitive units (MIRU) [84, 132, 133]. MIRU are direct tandem DNA repeats of 40-100 bp in size, identified in 41 different loci within intergenic regions of the MTC genome. The number of repeats of these interspersed in the MTC genome loci varies among different MTC strains. Out of the 41 loci, 12 were found to have sufficient polymorphism to be used as genetic markers [132, 135]; therefore, the MIRU-based genotype is a 12-digit number. The MIRU profiles were found to remain stable in vivo for at least 18 months [84]. The function of MIRU elements is not known yet; however, it was suggested that these human minisatellite-like VNTR regions may play a role in the evolution of the human host genome [134]. The discriminatory power of MIRU is comparable to that of the IS*6110*-RFLP method [84].

A single genotyping method does not posses enough discriminatory power to differentiate all unique isolates; therefore, two or more independent genetic markers should be used to achieve sufficient discrimination of MTC strains. The choice of the markers depends on the specific characteristics of MTC strains. Spoligotyping alone provides a good first-step discriminatory test, but in some cases should be followed by analyzes that are based on other genetic markers. In the cases when IS*6110* element disrupts DVRs, which results in their apparent loss, spoligotyping must be complemented by a secondary typing method to accurately asses the relationship of MTC strains [154]. Some spoligotype families, such as, for example, Beijing, are large, and their members can be discriminated further only by using other typing methods. The MIRU genotyping can also serve as a powerful first-line discrimination method [109]. Previous work showed that MIRU technique performs better than IS*6110*-RFLP when MTC strains have low copy numbers of IS*6110* [19, 124, 134].

Soini et al. [120] found that spoligotyping also helped discriminate this type of MTC strains.

2.2 Phylogeny of TB causing bacteria

For the purposes of molecular epidemiology, bacterial population genetic research aims at understanding the relationships between traits such as virulence, transmissibility, host specificity, and others, and mapping these traits onto the phylogenetic tree [50, 95].

MTC strains, as mentioned above, are characterized by atypically low degree of structural variation in the housekeeping genes and identical 16S rRNA sequences [127]. This led to the proposition that members of MTC appeared relatively recently [127]. Most MTC genetic variability is associated with insertion sequences, repetitive elements, and drug resistance phenotypes. However, MTC strains exhibit high phenotypic diversity. Also, their pathogenicity and host range differ significantly [12]. It was suggested that *M. tuberculosis (sensu stricto)* had evolved from *M. bovis*, which causes bovine TB, by specific adaptation of the animal pathogen to the human host [128]. This had been speculated before the sequencing of the whole genomes of MTC members was completed. When the complete genome of the *M. tuberculosis* clinical strain CDC1551 was sequenced [37] and compared to the whole-genome of the *M. tuberculosis* laboratory strain H37Rv [18] to identify polymorphic sequences with "potential relevance to disease pathogenesis, immunity, and evolution", large-sequence and single-nucleotide polymorphisms (SNPs) in numerous genes were discovered [37]. Results of another study proposed a new scenario on MTC evolution, showing that the common ancestor of the TB causing bacteria resembled *M. tuberculosis* or *M. canettii* and could have been a human pathogen [12]. This version placed *M. tuberculosis* closer to the common ancestor of MTC than *M. bovis* [12]. The completion of the *M. bovis* genome sequence confirmed this scenario [44].

Synonymous SNPs (sSNPs) are evolutionary neutral since they do not change the structure of proteins [49]. Thus, they are useful for large-scale studies revealing evolutionary relationships among bacterial strains, especially in strongly clonal species [50]. MTC strains have, on average, approximately one sSNP per 10,000 nucleotide sites [96]. Variation in two nonsynonymous (result in a change in translated amino acids) SNPs, at *katG* codon 463 (CTG or CGG) and at *gyrA* codon 95 (ACC to AGC), divided MTC strains into three principal genetic groups: ancestral group 1, and groups 2 and 3, where group 2 is ancestral to group 3 [127]. Eight clusters of related genotypes were identified in *M. tuberculosis* based on 148 sSNPs [50]. High-throughput sSNP genotyping has a potential to be used for rapid identification of clinical isolates. However, this method is dependent on availability of complete genome sequences of bacterial strains.

Several insertion sequences had been identified in MTC strains; they are reviewed elsewhere [93]. IS*6110* is the most widely used genetic marker to differentiate MTC strains. IS*6110*-RFLP-based phylogeny divided MTC strains into high-copy strains and low-copy strains [38], but genotyping using sSNP ruled out the idea that strains of *M. tuberculosis* with many IS*6110* copies and few IS*6110* copies are genetically distinct populations [50]. The variations in IS*6110*-RFLP fingerprints, rather than the IS*6110* copy number, are widely used for inferring relationships among MTC strains.

Spoligotype-based phylogeny of MTC strains has been recently derived [28, 123]. Eight independent genetic markers, IS*6110*, IS*1081*, the DR locus, and five VNTRs were used to infer phylogenetic relationships among a set of 90 clinical isolates of MTC bacteria [125]. However, integration of the information obtained from different genetic markers is not trivial [4]. To derive MTC strains phylogeny, the investigators used phylogenetic algorithms incorporated into different software packages, the most common of which were Taxotron (for example, [21]) and Bionumerics [122]. The Jaccard index [58], which does not account for the biological nature of spoligotypes but nevertheless has become popular in spoligotyping data analysis [121], was utilized as a means of pairwise comparison of spoligotypes. The authors assessed the performance of several phylogenetic tree construction methods on spoligotyping data [28]. In one of the most recent papers, combined spoligotyping/VNTR data were used to build a genetic network [122].

Although these studies provided useful insights into the global distribution and evolution of spoligotypes and other DNA fingerprints, they did not produce well defined phylogenetic relationships between MTC strains.

Most of the algorithms for phylogenetic trees are based on the assumption of marker independence. This is suitable, with some approximation, when MIRU, IS*6110*, or SNP data are analyzed, since these elements are distributed randomly in the chromosome and can be assumed to evolve independently. Results of previous work suggested that the deletion of contiguous DVR sequences did not occur sequentially, but rather by a single loss of several adjacent DVRs. This complicates the use of spoligotypes for deriving MTC phylogeny [154]. We believe that Dollo parsimony method or its modifications could potentially be suitable for deriving MTC phylogeny using spoligotype data, since it asserts that in evolution it is harder to gain a complex feature than to lose it [33]. This has not been investigated in this work and can be suggested as one of the future directions.

In summary, despite some success in using SNPs [4, 50, 127], spoligotyping [28], and spoligotyping/VNTR [122] data for deriving the phylogeny of MTC strains, the global picture still remains to be elucidated.

2.3 Data mining of MTC genotyping data: Current state

The first work on automatic classification of spoligotyping patterns described decision trees induced from the global spoligotyping database DB1 [113]. The spoligotyping patterns contained in the database were visually observed and manually labelled by a human expert, thus grouping the data into families. Most of the spoligotyping families of *M. tuberculosis* strains were named according to their prevalence in particular countries or regions. Some families were named by the place of their first discovery. For example, *M. tuberculosis* Haarlem family was initially identified in Dutch town, Haarlem [70]. Families such as *M. tuberculosis* LAM, where "LAM" stands for Latin American and Mediterranean, *M. tuberculosis* EAI (East African-Indian), and *M. tuberculosis* CAS (Central Asian), were named from the geographical areas where their members most commonly occur [125].

The C4.5 induction algorithm was employed to build the decision trees and produce "intelligible knowledge rules" from labelled spoligotyping data. Before this, a prototype selection algorithm was applied to eliminate "uninformative" examples. The authors claimed that one of the most significant contributions of their paper was in developing a method of automatic classification of MTC strains based on their spoligotype patterns using less than 43 spacers. They suggested eliminating the detection of the "uninformative" spacers from the laboratory setup. The spoligotype patterns were classified into nine major families that were called clades [113]. Another work of the same research group concentrated on using the decision tree approach to classify MIRU data for the validation of the families previously defined by spoligotyping [34].

The potential problem with the decision tree approach comes from the fact that the MTC phylogeny has yet to be resolved; therefore, the correctness of the manual labelling of the spoligotyping data is questionable. DB1 contained 342 spoligotypes and they all had to be labelled by a human expert. Moreover, the elimination of some of the spacers from the analysis does not seem reasonable, since the spoligotype is obtained from a single locus in the MTC chromosome. Interdependencies of some spacers were suggested, which indicates that we need to look at the spoligotyping pattern as a whole. Besides, spoligotyping is fast and inexpensive; therefore, detecting fewer spacers would not significantly decrease the cost of labor and laboratory materials.

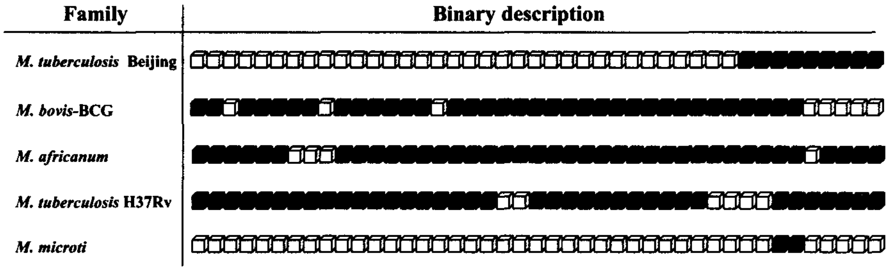


Figure 2.1: Examples of spoligotype family prototypes extracted using visual recognition rules from SpolDB3 database. Black cube indicates spacer, white cube indicates absence of spacer

Two highly cited papers described the current content of the global spoligotyping database SpolDB3 [35, 36]. Visual recognition rules were used to define 36 spoligotype families within the database [35]. Figure 2.1 gives an example of binary description of prototype spoligotypes for several such families. The authors also assessed the biogeographical specificity and the geographical spreading of spoligotype shared types (found in at least two patients) [36]. When this thesis was near completion, SpolDB4, an update of the global spoligotyping database, became publicly available [13]. We discuss SpolDB4 in Chapter 6.

In conclusion, epidemiological practice still lacks robust computational tools to analyze genotyping and epidemiological data. These tools should take into account biological characteristics of the data and be easily modified as new data become available.

CHAPTER 3
Mixture modeling for clustering

This chapter reviews mixture models and presents the Bernoulli mixture modeling approach based on the Naïve Bayes assumption. The expectation-maximization algorithm used to learn the parameters of the mixture models is described. Commonly used cluster validation techniques are discussed.

3.1 Clustering methods. Mixture models

Clustering or unsupervised classification is a process of grouping the objects in the data by some similarity measure. The clustering methods can be divided into two major categories: discriminative (distance-based methods, such as the well known k-means algorithm) and generative (model-based) methods [156]. Another common approach is to divide the clustering methods into partitional and hierarchical [60]. A partitional (flat) clustering divides the data samples into some number of groups. The classic example of this is the k-means algorithm. Hierarchical methods return a set of nested clusterings. These methods can either be agglomerative, when groups of data points are merged, or divisive, when at each step of the procedure the groups are divided.

Distance-based methods require a pairwise distance measure between data points, the most commonly used of which are Euclidean and Mahalanobis distances. Euclidean distance is the geometric distance in the multidimensional space and is computed as:

$$distance(x, y) = \left(\sum_i (x_i - y_i)^2\right)^{1/2}.$$

Calculating the distance between all pairs of data points is computationally inefficient and has a complexity of at least $O(N^2)$, where N is the number of data points. Moreover, it is often difficult to define a good distance metric, especially when dealing with complex data [156]. Similarity measures for complex data types, for example, biological sequences, are highly dependent on the data, require significant amount of expert knowledge, and in some cases are very difficult to formulate.

Model-based clustering techniques, such as mixture models, have the advantages that they do not require a distance metric. In addition, they are frequently more interpretable because the model for each cluster directly represents the cluster [156]. In the probabilistic context of density estimation, clustering can be viewed as "identifying the dense regions in the data source" [11]. The mixture model approach provides a useful and powerful framework for clustering data. The mixture model assumes that the data consist of some known or unknown number of component densities each of which corresponds to a cluster, or class [31, 87]. The probabilistic nature of the mixture models permits the use of different distributions that can handle complex data types. The most widely used model-based clustering method is the one based on learning a mixture of Gaussian distributions [6, 39, 155]. An important advantage of mixture models is that they handle uncertainty about cluster membership in a probabilistic manner allowing clusters overlap [119]. Each data point belongs to each cluster with some probability. Moreover, model-based methods are more amenable to incorporating prior knowledge than discriminative methods.

Clustering aims to answer the fundamental questions such as what is the underlying structure of the data, what clustering model best represents the data, and what is the "correct" number of clusters [107]. The problem of estimating mixture densities can be viewed as a missing data problem where the labels for the component densities are missing [46]. Some iterative procedure, most often the EM algorithm [24], is used to estimate parameters of the distribution. The most suitable model and the number of clusters can de determined by using some kind of a cluster validation technique [59].

3.2 Formal framework for Naïve Bayes assumption for mixture models

We begin by assuming that the multivariate Bernoulli mixture model generates the data and each model component satisfies the Naïve Bayes assumption.

Naïve Bayes is a well-known probabilistic classifier based on the assumption of feature independence. It has earned its popularity because of its simplicity, efficiency and robustness. Particularly suited for the high dimensionality problems, the Naïve Bayes performs surprisingly well compared to many more complicated classifiers that are not restricted by the independence assumption. The success of Naïve Bayes can be explained by the fact that misclassification error or zero-one loss is a function of the sign and does not necessarily reflect the quality of the fit to a probability distribution [25]. Rish and others discovered that Naïve Bayes achieves the best performance in two opposite cases: when the features are completely independent and when they are functionally dependent [104]. In an earlier work, it was shown that introducing attribute dependence does not necessarily improve the performance of Naïve Bayes [25]. Naïve Bayes uses Bayes' theorem [23]:

Theorem. *Let the events $A_1, \cdots, A_k$ form a partition of the space S such that $P(A_j) > 0$ for $j = 1, \cdots, k$, and let B be an event such that $P(B) > 0$. Then for $i = 1, \cdots, k$:*

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{k} P(A_j)P(B|A_j)}, \tag{3.1}$$

where $P(A_j|B)$ denotes the probability of event A_j given event B.

The events A are mutually exclusive, so formally we have:

$$P(B) = \sum_{A \in S} P(A|B). \tag{3.2}$$

From the definition of conditional probability, we have:

$$P(A, B) = P(B|A)(P(A). \tag{3.3}$$

Combining (3.2) and (3.3), we obtain (3.1), where $P(B|A_i)$ is a likelihood of B given A_i, $P(A_i)$ is a prior probability, and $P(A_i|B)$ is a posterior probability of A_i conditioned on B. Therefore, as we observe B, the prior probability $P(A_i)$ changes to posterior probability $P(A_i|B)$ [29].

Let X be a set of variables, or observations, which we want to classify. Each variable is a D-dimensional vector: $\mathbf{x} = \{x_1, \cdots, x_D\}$. Let C be a set of possible classes the variable can belong to: $C = \{c_1, \cdots, c_k\}$. Let's use the Bayes's rule to calculate the probability for each $\mathbf{x}$ to belong to c_j:

$$P(c_j|\mathbf{x}) = P(c_j)\frac{P(x_1, \cdots, x_D|c_j)}{P(x_1, \cdots, x_D)}, \; j = \{1, \cdots, k\}. \tag{3.4}$$

We assume the conditional independence of the features of each variable $\mathbf{x}$; therefore, $\mathbf{x}$ follows a multivariate distribution conditioned on class j:

$$P(x_1, \cdots, x_D|c_j) = P(\mathbf{x}|c_j) = \prod_{i=1}^{D} P(x_i|c_j), \; j = \{1, \cdots, k\}. \tag{3.5}$$

Using the Naïve Bayes assumption, the posterior probability of a class conditioned on an observed data point is defined as

$$P(c_j|\mathbf{x}) = P(c_j)\frac{\prod_{i=1}^{D} P(x_i|c_j)}{P(\mathbf{x})}, \; j = \{1, \cdots, k\}. \tag{3.6}$$

In practice, the denominator is a constant and the same for all classes; therefore, we are only interested in the numerator of the fraction.

The Naïve Bayes classifier assigns $\mathbf{x}$ to a class with the highest $P(c_j|\mathbf{x})$:

$$\nu_{NB} = \underset{c_j \in C}{argmax}\, P(c_j) \prod_{x_i \in \mathbf{x}} P(x_i|c_j). \tag{3.7}$$

We normalize each $P(c_j|\mathbf{x})$ in accordance with (3.4):

$$P(c_j|\mathbf{x}) = \frac{P(c_j)\prod_{i=1}^{d} P(x_i|c_j)}{\sum_{l=1}^{k} P(c_l)\prod_{i=1}^{d} P(x_i|c_l)}. \tag{3.8}$$

We apply Laplacian smoothing to avoid zero probabilities [85].

3.3 EM algorithm for Bernoulli mixture model

EM is a class of iterative algorithms for maximum likelihood (ML) or maximum a posteriori estimation useful for a variety of problems with incomplete or missing data [24]. The EM algorithm is an effective technique for estimating the parameters of the multivariate mixture model. On each iteration of the algorithm, there are two steps - the expectation (E) step and the maximization (M) step. In the E-step, the expected values of the missing data given the observed data and the current parameter estimates are computed to maximize the objective function of the algorithm, the total log-likelihood. In the M-step, the expected values of the missing data computed in the E-step are used to re-estimate the parameters. The steps are repeated until the difference between subsequent estimates is small. EM terminates at a locally optimal solution.

Let $X = \{x_1, \cdots, x_n\}$ be a collection of samples generated by a Bernoulli mixture model. Bernoulli distribution is simple, with two possible outcomes, "success" and "failure", that happen with probabilities p and $1-p$, respectively. Therefore, each of the 43 spoligotype positions is a Bernoulli distribution described by the parameter p. The whole spoligotyping pattern is modeled as a multivariate Bernoulli distribution. The mixture of these distributions represents the spoligotyping data. Each data point x_i is a D-dimensional vector $\{x_{id}, \cdots, x_{iD}\}$. Let's assume that the number, initial guesses of the mixing weights, and parameters of mixture components are known: $\Theta = (P(c_1), \cdots, P(c_k), \theta_1, \cdots, \theta_k)$, where $\theta_j = \{p_{j1}, \cdots, p_{jD}\}$. Each p_{jd} is the probability of spacer d being present given class c_j. The probability of spacer d being absent, given class c_j, is $1 - p_{jd}$. In accordance with the Naïve Bayes assumption of spacers' independence, we have:

$$P(x_i|c_j) = \prod_{d=1}^{D} p_{jd}^{x_{id}}(1-p_{jd})^{1-x_{id}}. \tag{3.9}$$

The log-likelihood function of parameters Θ, given data X, is defined as

$$L(\Theta|X) = \sum_{i=1}^{n} log \sum_{j=1}^{k} P(c_j)P(x_i|c_j). \tag{3.10}$$

According to the ML principle, a model best representing the data has parameters that maximize $L(\Theta|X)$. Since the ML problem cannot be solved in a closed form, the EM algorithm must be adopted to optimize the likelihood function. The EM algorithm iteratively refines an initial model to better fit the data. To be able to use EM, we need to introduce a hidden (missing) random variable $\mathbf{z}$ that indicates which component generated each data point. This means that with each data point x_i, we associate a vector variable $z_i = \{z_{i1}, \cdots, z_{ik}\}$ such that $z_{ij} = 1$ if x_i was generated by the j-th component and $z_{ij} = 0$ otherwise. The complete data log-likelihood function now can be expressed as follows:

$$L_c(\Theta|X, Z) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} logP(c_j)P(x_i|c_j). \tag{3.11}$$

L_c cannot be employed directly since $\mathbf{z}$ is unknown; therefore, according to the classical EM approach [87], we will work with the expectation of L_c, $Q(\Theta, \Theta')$. This expectation serves as a lower bound on the observed data likelihood:

$$Q(\Theta, \Theta') = \sum_{z} log(P(X, \mathbf{z}|\Theta))p(\mathbf{z}|X, \Theta'). \tag{3.12}$$

It was shown that the maximization of the function $Q(\Theta, \Theta')$ with respect to Θ is equivalent to maximizing the observing likelihood function in Eq. 3.10 [24]. Therefore, introduction of the indicator variable $\mathbf{z}$ allows us to decouple the maximization problem into a set of simple maximizations. Substituting Eq. 3.10 into the definition of the Q function, we obtain:

$$Q(\Theta, \Theta') = \sum_{i=1}^{n}\sum_{j=1}^{k} E[z_{ij}] logP(c_j)P(x_i|c_j). \tag{3.13}$$

The expected values of the hidden variables z_{ij} are defined as

$$E[z_{ij}] = \sum_{z} z_{ij}p(\mathbf{z}|X, \Theta') = \frac{P(c_j)'P(x_i|c_j)}{\sum_{l=1}^{k} P(c_l)'P(x_i|c_l)},\ i = \{1, \cdots, n\},\ j = \{1, \cdots, k\}, \tag{3.14}$$

where Θ' is the previous estimate of the parameters.

From Eq. 3.9, $E[z_{ij}]$ is estimated in the E-step of EM as follows:

$$E[z_{ij}] = \frac{P(c_j)' \prod_{d=1}^{D} p_{jd}^{x_{id}}(1 - p_{jd})^{1-x_{id}}}{\sum_{l=1}^{k} P(c_l)' \prod_{d=1}^{D} p_{ld}^{x_{id}}(1 - p_{ld})^{1-x_{id}}}, i = \{1, \cdots, n\}, j = \{1, \cdots, k\}. \tag{3.15}$$

The Q function (Eq. 3.13) is maximized in the M-step of the EM algorithm with respect to parameters Θ, given the estimated in the E-step expected values of hidden variables $E[z_{ij}]$:

$$\Theta = \underset{\Theta}{argmax} \sum_{i=1}^{n} \sum_{j=1}^{k} (E[z_{ij}] log P(c_j) + E[z_{ij}] log P(x_i|c_j)). \tag{3.16}$$

Each of the two terms on the right hand side can be optimized separately. To find the estimate for parameters $P(c_j)$, we use the Lagrangian multiplier and the constraint $\sum_{j=1}^{k} P(c_j) = 1$, thus obtaining the expression:

$$\frac{\partial Q(\Theta, \Theta')}{\partial P(c_j)} = \frac{\partial}{\partial P(c_j)} \left(\sum_{i=1}^{n} \sum_{j=1}^{k} E[z_{ij}] log P(c_j) + \lambda \left(\sum_{j=1}^{k} P(c_j) - 1 \right) \right) = 0. \tag{3.17}$$

Solving Eq. 3.17, we obtain:

$$P(c_j) = \frac{\sum_{i=1}^{n} E[z_{ij}]}{\sum_{i=1}^{n} \sum_{j=1}^{k} E[z_{ij}]} = \frac{\sum_{i=1}^{n} E[z_{ij}]}{n}, \; j = \{1, \cdots, k\}. \tag{3.18}$$

To obtain the optimizing value for p_{jd}, we use the following expression:

$$\frac{\partial Q(\Theta, \Theta')}{\partial p_{jd}} = \frac{\partial}{\partial p_{jd}} \left(\sum_{i=1}^{n} \sum_{j=1}^{k} E[z_{ij}] log \prod_{d=1}^{D} p_{jd}^{x_{id}} (1 - p_{jd})^{1 - x_{id}} \right) = 0. \tag{3.19}$$

To solve Eq. 3.19, we express it in a simpler form:

$$\frac{\partial}{\partial p_{jd}} \left(\sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{d=1}^{D} E[z_{ij}] log \, p_{jd}^{x_{id}} (1 - p_{jd})^{1 - x_{id}} \right) = 0, \tag{3.20}$$

from which we obtain the following:

$$\sum_{i=1}^{n} \frac{E[z_{ij}](x_{id} - p_{jd})}{p_{jd}(1 - p_{jd})} = 0. \tag{3.21}$$

Solution of the equation above gives us the estimate:

$$p_{jd} = \frac{\sum_{i=1}^{n} E[z_{ij}] x_{id}}{\sum_{i=1}^{n} E[z_{ij}]}. \tag{3.22}$$

Algorithm 1 presents a concise summary of the EM algorithm for the multivariate Bernoulli mixture model.

Algorithm 1: EM algorithm for multivariate Bernoulli mixture model.

```
1. Choose initial parameter settings $\Theta' = \{P'(c_1), \cdots, P'(c_k), \theta'_1, \cdots, \theta'_k\}$.
   2. Repeat until convergence:
      E-step:
      $*E[z_{ij}] = \frac{P'(c_j)\prod_{d=1}^{D} p_{jd}^{x_{id}}(1-p_{jd})^{1-x_{id}}}{\sum_{l=1}^{k} P'(c_l)\prod_{d=1}^{D} p_{ld}^{x_{id}}(1-p_{ld})^{1-x_{id}}}, i = \{1, \cdots, n\}, j = \{1, \cdots, k\}$.

      M-step:
      $*P(c_j) = \frac{\sum_{i=1}^{n} E[z_{ij}]}{n}, j = \{1, \cdots, k\}$.
      $*p_{jd} = \frac{\sum_{i=1}^{n} E[z_{ij}]x_{id}}{\sum_{i=1}^{n} E[z_{ij}]}, j = \{1, \cdots, k\}, d = \{1, \cdots, D\}$.
      *Set $\Theta = \{P(c_1), \cdots, P(c_k), \theta_1, \cdots, \theta_k\}$.
```

The probability estimates for p_{jd} can be smoothed using a Laplacian prior [85]:

$$p_{jd} = \frac{\sum_{i=1}^{n} E[z_{ij}]x_{id} + 1}{\sum_{i=1}^{n} E[z_{ij}] + 2}, \; j = \{1, \cdots, k\}, \; d = \{1, \cdots, D\}. \tag{3.23}$$

Iterating between the E- and M-steps results in nondecreasing sequence of values for the total log-likelihood. We should avoid starting the EM algorithm from a "pathological" point, where p_{jd} parameters for all components of a mixture model are the same. If EM is initialized from a non-pathological stating point, the algorithm will always achieve a proper stationary point of the log-likelihood [16].

Initialization of the EM algorithm requires special attention, since the solution is highly dependent upon its starting points. The initialization includes choosing both the most appropriate parameter values and the number of components in the mixture model.

The next section discusses how to address the problem of assessing cluster quality in order to select the most appropriate model for the given data.

3.4 Cluster quality assurance

In general, assessing the clustering results and interpreting the clusters found are as important as generating the clusters [59]. Unifying model assessment procedures have not been developed in the clustering area [75]. Cluster validation techniques can be based on an external criterion, which is some type of information not contained in the data set. The cluster quality can also be evaluated by some internal criteria that stem exclusively from the data [51, 107].

The least computationally demanding and the most practical model selection methods are based on maximizing the log-likelihood of the data. The log-likelihood increases with increasing the number of mixture components; therefore, penalizing criteria are used to account for model complexity. The most common of these are the information-theoretic criteria such as the Akaike Information Criterion (AIC) [2], Bayesian Information Criterion (BIC) [112], and the Minimum Description Length (MDL) [98, 105]. These criteria allow comparison of models with different parameter settings and/or different number of components [40]. However, it has been shown that the log-likelihood dominates the penalty terms in the AIC and MDL, therefore making them useless for selecting the number of components in the model [74]. The fully Bayesian approach treats the number of clusters as a random variable and calculates a posterior distribution on this variable given the data and the model. This procedure is computationally cumbersome; even for the most widely used Gaussian mixture model, there is no closed form solution for the posterior [119].

Other methods include bootstrap [30] and cross-validation techniques, widely used in supervised learning. They have not been used much in clustering, perhaps because for many clustering techniques "there is no obvious score-function to cross-validate" [119]. This problem does not exist for probabilistic clustering because any score criterion that measures how well the model fits the data can be used for the model selection. Cross-validation consists of randomly dividing the data set into train and test set, fitting the model to the train partition and then testing it on the test set. Different cross-validation methodologies exist differing in how the partitions are chosen. The common procedures are the "v-fold" cross-validation and the "leave-one-out" cross-validation. Another specific cross-validation method is based on dividing the data set into M partitions. Each of these partitions is treated as a test set, which is a β fraction of the data set, and the train set, which is a $1 - \beta$ fraction of the data set. This approach was termed "Monte Carlo cross-validation" (MCCV) [115]. The test and train sets do not have to be disjoint [118], but the most common approach is to make them non-overlapping.

Smyth (1996) investigated the use of cross-validated by MCCV approach log-likelihood in model selection, particularly in choosing the number of components in a mixture model [118]. His results showed that this procedure, with $\beta = 0.5$, performed better than the 10-fold cross-validation. The problem of selecting the number of clusters in the data is also called model order selection. There can be cases where the number of clusters is known, but in the majority of the situations we need to determine it from the data. Several methods have been proposed to find the most appropriate number of clusters in the data. For example, Tibshirani et al. (2001) proposed a method called "Gap statistics", which uses the Euclidean distance measure [139].

The resampling-based approach requires no prior information and seeks to assess the cluster stability to find the most self-consistent data partitioning [107]. In this approach, the data set is iteratively split into two sets, both of which are used as train sets, and the disagreement between clustering solutions for these sets is computed [75]. The splitting process either picks the randomly selected data points or employs random parameters generated from the data [52].

The model initialization consists in selecting the starting parameters for an iterative procedure, which then refines the initial settings to fit the data. The EM algorithm, widely used for the estimation of the parameters for the mixture model, usually converges to a local maximum of the log-likelihood function. This hill-climbing procedure strongly depends on the starting points. The algorithm is usually restarted several times and the highest log-likelihood solution is used. Agglomerative hierarchical clustering [40], k-means clustering [52] and other techniques may also be used to determine initial parameters. Useful in the context of this study, the analysis of the EM initialization techniques for estimating Bernoulli mixture models was recently performed [64]. In another recent work, the cluster initialization methods were divided into three major families, such as random sampling, distance optimization methods, and density estimation methods [52].

A special group of methods which inherently test for cluster stability is clustering ensembles. They go beyond what is achieved by a single clustering algorithm and are characterized by robustness, novelty, confidence estimation, and parallelization and scalability [142]. Various clustering algorithms produce multiple different data partitions, which are combined in a consensus partition that ideally should have a better quality than the given partitions [143]. The problem of combining multiple clusterings into a single solution has no less challenges than single clustering methods. The choice of the clustering algorithm(s), number of clusters in a consensus partition, and particularly an algorithm to deduce the partition, all constitute the challenges.

Co-association matrix as a means to summarize the results of multiple clusterings was suggested. The final clustering was determined using a voting k-means algorithm [41]. In hypergraph methods, the clusters resulted from multiple clusterings are represented as hyperedges on a graph, where each hyperedge describes a set of objects belonging to the same cluster. The final clustering is found using a k-way min-cut hypergraph partitioning problem [129]. Another approach is based on using generalized mutual information [141]. Among several other interesting methods, a probabilistic modeling of consensus partition can be mentioned [142]. This method, however, leaves us with the problem of determining the number of clusters in the consensus.

Ultimately, irrespective of the cluster validation technique that we use to decide on the "best" model, we often still need to visually explore the results of several best-fitting models and, using as much domain knowledge as possible, select the most suitable one.

CHAPTER 4
SPOTCLUST

In this section, we present the results of the application of multivariate Bernoulli mixture models to the spoligotyping data obtained from the TB Control Center of the Public Health Research Institute, and from the Division of Infectious Diseases of the Wadsworth Center. SPOTCLUST (SPOligotyping data on Tuberculosis CLUSTering) is a first step in an ongoing project aimed at developing mathematical models for different genotyping and epidemiological data on TB and other infectious diseases.

4.1 Spoligotyping data

We applied the algorithm to 8011 spoligotype patterns obtained from MTC strains isolated between 1996 and 2004, primarily from New York State TB patients. Out of these patterns, 535, identified among 7166 MTC strains, each represented a shared type, i.e. observed in at least two specimens obtained from different patients, and 845 were unique isolates or orphans, each recovered from only one patient.

4.2 Probabilistic framework

The model-based approach is the most appropriate for spoligotype data clustering, because the distance measure between spoligotypes has not been determined yet. Moreover, this approach allows us to incorporate the prior knowledge on the evolution of spoligotype patterns. The probabilistic framework that we adopted assumes that a multivariate Bernoulli mixture model generates the data and that there is a one-to-one correspondence between mixture model components and spoligotype families [31]. Let X be our database of n spoligotypes. The goal is to label each spoligotype. Each spoligotype is a binary 43-dimensional vector: $\mathbf{x} = \{x_1, \cdots, x_{43}\}$. Let C be a mixture model, which is a set of components: $C = \{c_1, \cdots, c_k\}$. Each mixture component $c_j \in C$ is defined by a set of parameters, which are the mixing weight of the component, $P(c_j)$, and a 43-variable Bernoulli distribution, θ_j. The mixing weights satisfy the constraints:

$$\sum_{j=1}^{k} P(c_j) = 1 \text{ and } P(c_j) \geq 0. \tag{4.1}$$

The probability of a spoligotype $\mathbf{x}$ being generated by a model C is

$$P(\mathbf{x}) = \sum_{j=1}^{k} P(c_j)P(\mathbf{x}|c_j). \tag{4.2}$$

Thus, to generate a spoligotype, first a mixture component is chosen with a probability $P(c_j)$, then its parameters are used to produce a binary spoligotype sequence. Let us denote each variable of spoligotype $\mathbf{x}_i$ as x_{id}, $i = \{1, \cdots, n\}$. Each mixture component c_j has 43 parameters p_{jd}, where each p_{jd} is a probability of a spacer being present and $1 - p_{jd}$ is a probability of a spacer being absent at a position d of a spoligotype. The probability of a spoligotype $\mathbf{x}_i$ given component c_j is

$$P(\mathbf{x}_i|c_j) = \prod_{d=1}^{43} p_{jd}^{x_{id}}(1 - p_{jd})^{1-x_{id}}, \; j = \{1, \cdots, k\}. \tag{4.3}$$

The parameters for finite mixture models are often estimated by the ML approach. The EM algorithm, described above, is the most commonly used algorithm for finding a local maximum of the observed data likelihood function.

4.3 Multivariate Bernoulli model with Hidden Parent

A widely accepted hypothesis states that spoligotypes evolve by the deletion of a single or multiple contiguous spacers and that the spacer duplication is a very rare event [3, 32, 148, 154]. To accurately model spoligotypes, the multivariate Bernoulli model was modified to incorporate this knowledge in the form of a "Hidden Parent". Given a 43-dimensional spoligotype $\mathbf{x}_i$ and a spacer position d, if $x_{id} = 1$ (spacer present), then the distribution c_j generating $\mathbf{x}_i$ should have the probability p_{jd} very high. In other words, we have assumed that each spoligotype family has an unobserved Hidden Parent and that the children of the Parent, the observed strains in this family, may lose a spacer with some small probability, but are extremely unlikely to gain one. If we observe $x_{id} = 0$ in the spoligotype, then its Hidden Parent should be generating a 0 with high probability and a 1 with some non-negligible probability (the child can lose a spacer) at position d.

Given the i^{th} spoligotype and the d^{th} spacer position, we introduce the following notations: $m_{11} = P(x_{id} = 1|H_d = 1)$, $m_{00} = P(x_{id} = 0|H_d = 0)$, $m_{10} = P(x_{id} = 1|H_d = 0)$, and $m_{01} = P(x_{id} = 0|H_d = 1)$, where H_d is the d^{th} spacer in the Hidden Parent. It is obvious that $m_{01} = 1 - m_{11}$ and $m_{00} = 1 - m_{10}$; therefore, we can work with only two parameters, m_{11} and m_{10}. Since no prior knowledge exists on the probabilities with which the child spoligotype loses spacers from its parent spoligotyping pattern, we assume that these parameters are the same for each mixture component and each spacer position. We assume that the probabilities of the child gaining and losing spacers from the parent's pattern are 10^{-7} and 10^{-1}, respectively; therefore, $m_{11} = 0.9$ and $m_{10} = 10^{-7}$.

Equation 4.3, defining the probability of spoligotype $\mathbf{x}_i$ given mixture component c_j, becomes:

$$P(\mathbf{x}_i|c_j) = \prod_{d=1}^{43}(p_{jd}m_{11}+(1-p_{jd})m_{10})^{x_{id}}((1-p_{jd})(1-m_{10})+p_{jd}(1-m_{11}))^{1-x_{id}}. \quad (4.4)$$

The log-likelihood function of Θ given the data X is now defined in the following form:

$$L(\Theta|X) =$$

$$\sum_{i=1}^{n} log \sum_{j=1}^{k} P(c_j) \prod_{d=1}^{43}(p_{jd}m_{11} + (1-p_{jd})m_{10})^{x_{id}}((1-p_{jd})(1-m_{10}) + p_{jd}(1-m_{11}))^{1-x_{id}}. \quad (4.5)$$

Since we use the EM algorithm to estimate the parameters p_{jd}, we introduce, as described before, a missing random variable $\mathbf{z}$ that indicates which component of the mixture model generated each data point. $E[z_{ij}]$ is now defined as

$$\frac{P(c_j)' \prod_{d=1}^{43}(p_{jd}m_{11} + (1-p_{jd})m_{10})^{x_{id}}((1-p_{jd})(1-m_{10}) + p_{jd}(1-m_{11}))^{1-x_{id}}}{\sum_{l=1}^{k} P(c_l)' \prod_{d=1}^{43}(p_{ld}m_{11} + (1-p_{ld})m_{10})^{x_{id}}((1-p_{ld})(1-m_{10}) + p_{ld}(1-m_{11}))^{1-x_{id}}}, \quad (4.6)$$

where $i = \{1, \cdots, n\}, j = \{1, \cdots, k\}$.

Analogously to the Eq. 3.19, the expression for the optimizing value for p_{jd} is defined as

$$\frac{\partial}{\partial p_{jd}} \sum_{i=1}^{n} \sum_{j=1}^{k} E[z_{ij}] log \prod_{d=1}^{43} (p_{jd}m_{11}+(1-p_{jd})m_{10})^{x_{id}}((1-p_{jd})(1-m_{10})+p_{jd}(1-m_{11}))^{1-x_{id}}, \tag{4.7}$$

which can be expressed in a simpler form:

$$\frac{\partial}{\partial p_{jd}} \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{d=1}^{43} E[z_{ij}] log (p_{jd}m_{11}+(1-p_{jd})m_{10})^{x_{id}}((1-p_{jd})(1-m_{10})+p_{jd}(1-m_{11}))^{1-x_{id}}. \tag{4.8}$$

Taking the derivative of the Eq. 4.8 and setting the result to zero yield the following:

$$\sum_{i=1}^{n} \frac{E[z_{ij}](m_{11}-m_{10})(m_{10}-x_{id}+p_{jd}(m_{11}-m_{10}))}{(m_{11}p_{jd}+m_{10}(1-p_{jd}))(m_{10}-1+p_{jd}(m_{11}-m_{10}))} = 0. \tag{4.9}$$

Solving this equation, we obtain:

$$p_{jd} = \frac{\sum_{i=1}^{n} E[z_{ij}]x_{id} - \sum_{i=1}^{n} E[z_{ij}]m_{10}}{\sum_{i=1}^{n} E[z_{ij}](m_{11}-m_{10})}. \tag{4.10}$$

The optimizing values for mixing weights of model components are obtained in the same way as described above (see Eq. 3.17 and Eq. 3.18).

Algorithm 2 outlines EM for the multivariate Bernoulli mixture model with Hidden Parent.

We will discuss shortly that we have analyzed mostly the shared types, i.e. identical spoligotypes characterizing isolates obtained from two or more patients. The number of occurrences of each spoligotype pattern was important for the clustering task. However, repeatedly processing identical data points among the total of more than 7000 spoligotypes is redundant and significantly slows the EM algorithm. We improved this by exploiting each shared type only once and accounting for the number of times it is found in the database. If we let n be a total number of distinct shared types and denote as t_i the number of occurrences of the shared type i, then

Algorithm 2: EM algorithm for multivariate Bernoulli mixture model with Hidden Parent.

```
1. Choose initial parameter setting $\Theta' = \{P'(c_1), \cdots, P'(c_k), \theta'_1, \cdots, \theta_k\}$.
  2. Repeat until convergence:
     E-step:
     $*E[z_{ij}] = \frac{P'(c_j)\prod_{d=1}^{43}(p_{jd}m_{11}+(1-p_{jd})m_{10})^{x_{id}}((1-p_{jd})(1-m_{10})+p_{jd}(1-m_{11}))^{1-x_{id}}}{\sum_{l=1}^{k}P'(c_l)\prod_{d=1}^{43}(p_{ld}m_{11}+(1-p_{ld})m_{10})^{x_{id}}((1-p_{ld})(1-m_{10})+p_{ld}(1-m_{11}))^{1-x_{id}}}$.

     M-step:
     $*P(c_j) = \frac{\sum_{i=1}^{n}E[z_{ij}]}{n}, j = \{1, \cdots, k\}$.
     $*p_{jd} = \frac{\sum_{i=1}^{n}E[z_{ij}]x_{id}-\sum_{i=1}^{n}E[z_{ij}]m_{10}}{\sum_{i=1}^{n}E[z_{ij}](m_{11}-m_{10})}, j = \{1, \cdots, k\}, d = \{1, \cdots, 43\}$.
     $*\text{Set } \Theta = \{P(c_1), \cdots, P(c_k), \theta_1, \cdots, \theta_k\}$.
```

at the M-step of EM, we calculate mixture component weights as follows:

$$*P(c_j) = \frac{\sum_{i=1}^{n} E[z_{ij}]t_i}{N}, \ j = \{1, \cdots, k\}, \tag{4.11}$$

where N is the total number of the data points.

The log-likelihood function of Θ is expressed in the following form:

$$L(\Theta|X) = \sum_{i=1}^{n}(log\sum_{j=1}^{k}P(c_j)P(x_i|c_j))t_i. \tag{4.12}$$

In the text below, for simplicity of representation, we omit the fact that the shared types from the New York State database were each used once in the EM algorithm.

4.4 Model initialization and validation

The performance of the method is highly dependent on the initialization (seeding) of EM, which includes choosing the number of components in a mixture model and their parameters. To incorporate expert knowledge, we used the prototypes derived from SpolDB3. Figure 4.1 shows these prototypes as well as expert defined visual recognition rules [35]. We extracted seeds for 32 mixture components. The prototypes for the *M. africanum* and *M. tuberculosis* subfamily CAS were combined into seeds for two corresponding components; and the prototype for *M. canetti* was excluded from the analysis. Based on visual inspection, we supplemented these seeds with four additional prototypes for spoligotypes that did not match any of the SpolDB3-based prototypes.

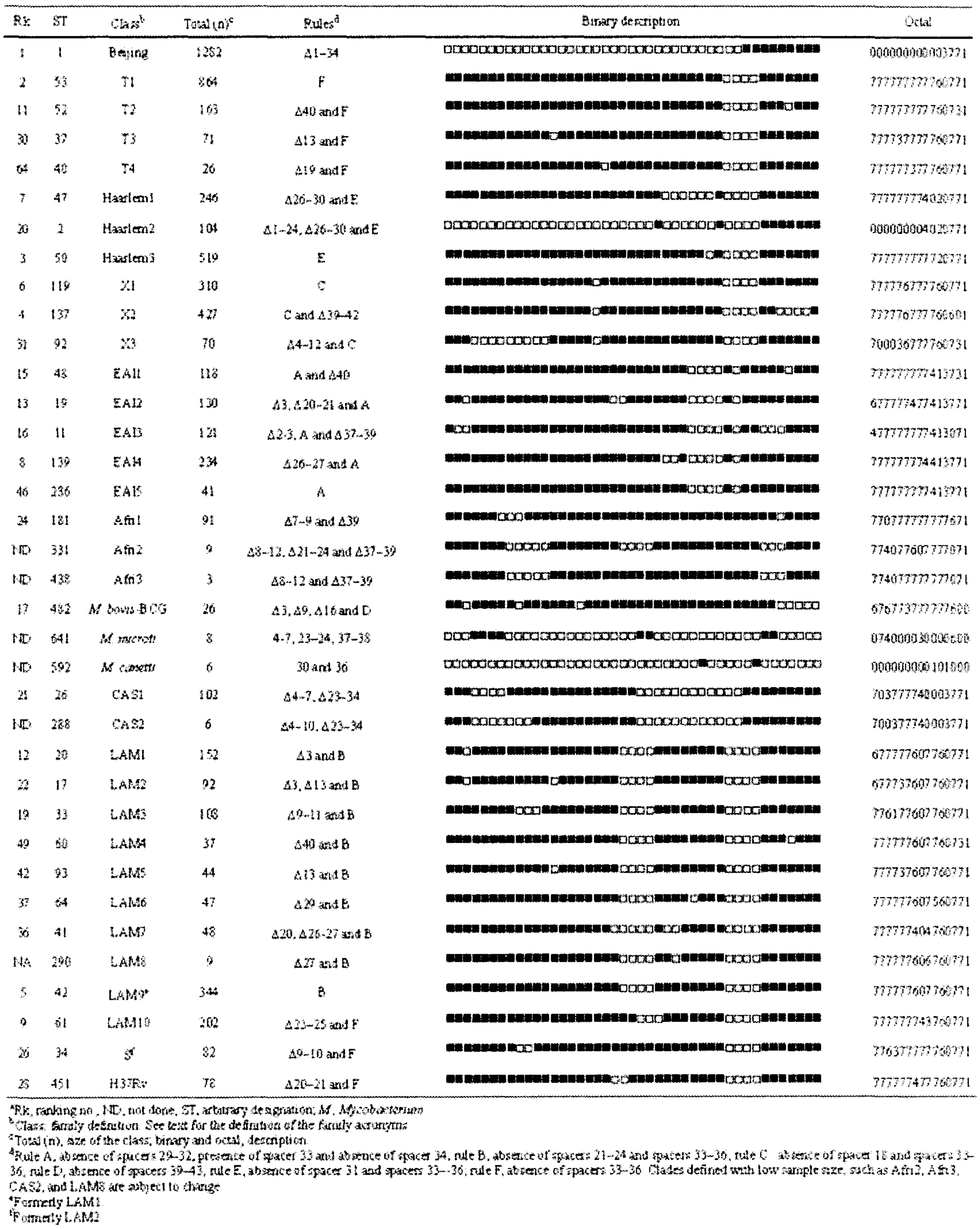
In another model, EM was initialized randomly. We employed Monte Carlo cross-validation (MCCV) approach [118] to find k, the number of components in the mixture. MCCV divides the data M times randomly into disjoint test and train partitions. Smyth (1996) samples his data set with replacement [118], but the most accepted opinion states that common points in test and train sets can potentially increase the stability artificially [75]. The test subset is fraction β of the data set. For each of the M partitions, we vary k from k_{min} to k_{max}. EM is initialized using the k-means algorithm, which is itself initialized randomly. Figure 4.2 outlines the MCCV approach. EM is randomly restarted 10 times and the highest log-likelihood solution is used as a trained model. The random initialization is schematically depicted in Fig. 4.3.

EM iterates until the total log-likelihood change is less then 10^{-7} or until the change of component weights' sum is less than 5×10^{-8}. Alternatively, it stops when the number of iterations reaches 30. For the highest-total-log-likelihood model, the EM algorithm iterates 300 additional times or until convergence. Before starting the 300 iterations, each prototype p_{jd} is modified by adding randomness component to it [64]:

$$p_{jd} = \alpha p_{jd}^{rand} + (1 - \alpha)(p_{jd}), \tag{4.13}$$

where α, $p_{jd}^{rand} \in (0, 1)$ and α measures the "global randomness" of p_{jd}.

Each trained k-order model is applied to the test set, and the test data log-likelihood is calculated. The procedure is repeated M times and the average (cross-validated) test data log-likelihood, $\hat{L}_k^{cv}$, is calculated for each k. It has been shown that $\hat{L}_k^{cv}$ is "an approximately unbiased estimator" of the expected value of the Kullback-Leibler (KL) distance [71, 72] between the real and the estimated data-generating probability distribution [118]. The KL distance, or divergence, also called relative entropy, between the "true" discrete distribution having probability function p_k and an estimated, or "target", discrete distribution having probability function



Rk	ST	Class[b]	Total (n)[c]	Rules[d]	Binary description	Octal
1	1	Beijing	1282	Δ1–34	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■	000000000003771
2	53	T1	864	F	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777777777760771
11	52	T2	103	Δ40 and F	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■	777777777760731
30	37	T3	71	Δ13 and F	■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	777737777760771
64	40	T4	26	Δ19 and F	■■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	777777377760771
7	47	Haarlem1	246	Δ26–30 and E	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	777777774020771
20	2	Haarlem2	104	Δ1–24, Δ26–30 and E	□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□□■□□□□■■■■■■■	000000004020771
3	50	Haarlem3	519	E	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■	777777777720771
6	119	X1	310	C	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	777776777760771
4	137	X2	427	C and Δ39–42	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■□□□□■	777776777760601
31	92	X3	70	Δ4–12 and C	■■■□□□□□□□□□□■■■■□■■■■■■■■■■■■■■□□□□■■■□■■■	700036777760731
15	48	EAI1	118	A and Δ40	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■□■■■	777777777413731
13	19	EAI2	130	Δ3, Δ20–21 and A	■■□■■■■■■■■■■■■■■■■□□■■■■■■■□□□□■□■■■■■■■■■	677777477413771
16	11	EAI3	121	Δ2-3, A and Δ37–39	■□□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■□□□■■■■	477777777413071
8	139	EAI4	234	Δ26–27 and A	■■■■■■■■■■■■■■■■■■■■■■■■■□□■□□□□■□■■■■■■■■■	777777774413771
46	236	EAI5	41	A	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■■■■■	777777777413771
24	181	Afri1	91	Δ7–9 and Δ39	■■■■■■□□□■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■	770777777777671
ND	331	Afri2	9	Δ8–12, Δ21–24 and Δ37–39	■■■■■■■□□□□□■■■■■■■■□□□□■■■■■■■■■■■■□□□■■■■	774077607777071
ND	438	Afri3	3	Δ8–12 and Δ37–39	■■■■■■■□□□□□■■■■■■■■■■■■■■■■■■■■■■■■□□□■■■■	774077777777071
17	482	*M. bovis*-BCG	26	Δ3, Δ9, Δ16 and D	■■□■■■■■□■■■■■■□■■■■■■■■■■■■■■■■■■■■■■□□□□□	676773777777600
ND	641	*M. microti*	8	4-7, 23–24, 37–38	□□□■■■■□□□□□□□□□□□□□□□■■□□□□□□□□□□□□■■□□□□□	074000030000600
ND	592	*M. canetti*	6	30 and 36	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□■□□□□□□□	000000000101000
21	26	CAS1	102	Δ4–7, Δ23–34	■■■□□□□■■■■■■■■■■■■■■■□□□□□□□□□□□□■■■■■■■■■	703777740003771
ND	288	CAS2	6	Δ4–10, Δ23–34	■■■□□□□□□□■■■■■■■■■■■■□□□□□□□□□□□□■■■■■■■■■	700377740003771
12	20	LAM1	152	Δ3 and B	■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	677777607760771
22	17	LAM2	92	Δ3, Δ13 and B	■■□■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	677737607760771
19	33	LAM3	108	Δ9–11 and B	■■■■■■■■□□□■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	776177607760771
49	60	LAM4	37	Δ40 and B	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■□■■■	777777607760731
42	93	LAM5	44	Δ13 and B	■■■■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	777737607760771
37	64	LAM6	47	Δ29 and B	■■■■■■■■■■■■■■■■■■■■□□□□■■■■□■■■□□□□■■■■■■■	777777607560771
36	41	LAM7	48	Δ20, Δ26-27 and B	■■■■■■■■■■■■■■■■■■■□□□□□■□□■■■■■□□□□■■■■■■■	777777404760771
NA	290	LAM8	9	Δ27 and B	■■■■■■■■■■■■■■■■■■■■□□□□■■□■■■■■□□□□■■■■■■■	777777606760771
5	42	LAM9[e]	344	B	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	777777607760771
9	61	LAM10	202	Δ23–25 and F	■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■■■□□□□■■■■■■■	777777743760771
26	34	S[f]	82	Δ9–10 and F	■■■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	776377777760771
25	451	H37Rv	78	Δ20–21 and F	■■■■■■■■■■■■■■■■■■■□□■■■■■■■■□□□□■■■■■■■	777777477760771

[a]Rk, ranking no., ND, not done, ST, arbitrary designation; *M.*, *Mycobacterium*
[b]Class: family definition. *See text for the definition of the family acronyms*
[c]Total (n), size of the class; binary and octal, description.
[d]Rule A, absence of spacers 29–32, presence of spacer 33 and absence of spacer 34, rule B, absence of spacers 21–24 and spacers 33–36, rule C absence of spacer 18 and spacers 33–36, rule D, absence of spacers 39–43, rule E, absence of spacer 31 and spacers 33–-36; rule F, absence of spacers 33–36 Clades defined with low sample size, such as Afri2, Afri3, CAS2, and LAM8 are subject to change
[e]Formerly LAM1
[f]Formerly LAM2

Figure 4.1: Excerpt from SpolDB3 database showing phototypes, visual recognition rules, and binary and octal description [35]

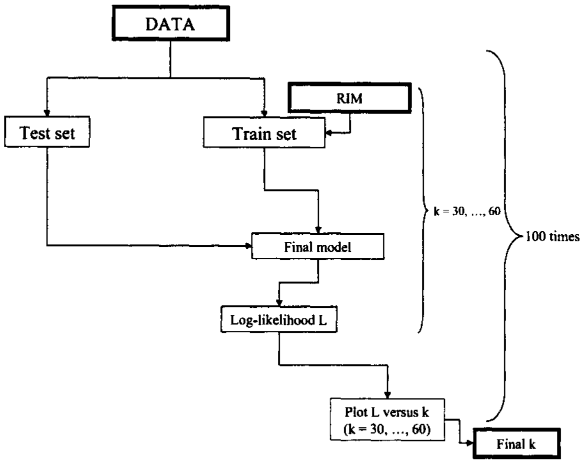


Figure 4.2: Schema of the MCCV approach used to find the optimal number of clusters

q_k is defined by:

$$D(p_k, q_k) = \sum_k p_k log \frac{p_k}{q_k}.$$

For continuous distributions, the summation is replaced by the integral. This is not a true metric, since the distance is not symmetric: $D(p_k, q_k) \neq D(q_k, p_k)$. Nevertheless, it possesses some useful properties and is widely used in information theory and probability theory as a natural measure of distance from a true distribution to some other arbitrary distribution. Computation of the KL divergence for Gaussian mixture model is not direct [92]. The same applies to the multivariate Bernoulli mixture model.

The plot of $\hat{L}_k^{cv}$ as a function of k shows what k is the most probable for the given data. Our algorithm was run with $M = 100$, $\beta = 0.3$, $\alpha = 0.7$, and k varying from 30 to 60.

After we had decided on a particular k, we generated, as previously described (see Fig. 4.3), 100 randomly initialized mixture models and calculated the total stabilities (over the resulting families) for each of them relative to the other 99 models. We chose a final mixture model based on the total stability, or average best match [55], and the total log-likelihood (see Fig. 4.4). We call the stability

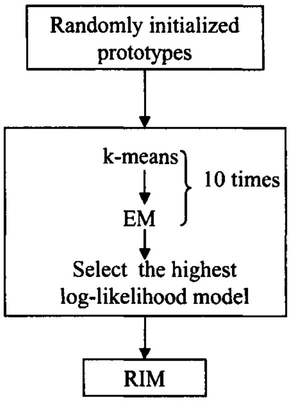

Randomly initialized prototypes
k-means
EM
10 times
Select the highest log-likelihood model
RIM

Figure 4.3: Schema of the random initialization of the EM algorithm

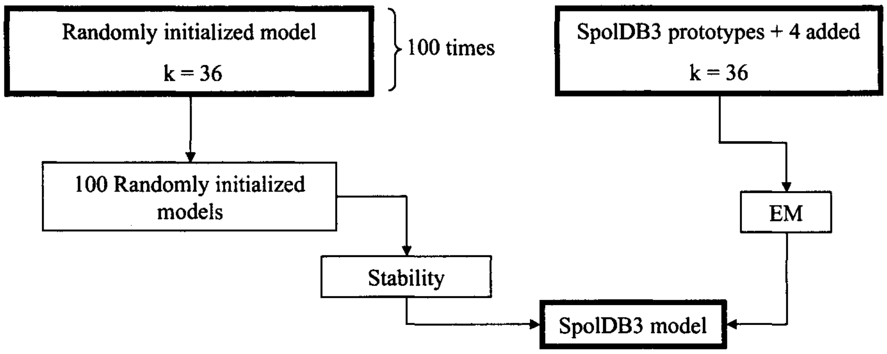

Randomly initialized model
k = 36
100 times
SpolDB3 prototypes + 4 added
k = 36
100 Randomly initialized models
EM
Stability
SpolDB3 model

Figure 4.4: Schema of the approach used to find the probabilistically best model

of a spoligotype cluster, or family, the average best match between this cluster and clusters identified using other models. If we define two clusters C and C' and treat them as sets, the match (between 0 and 1) will be defined as [55]:

$$match(C, C') = min(\frac{|C \bigcap C'|}{|C|}, \frac{|C \bigcap C'|}{|C'|}). \tag{4.14}$$

High match values mean that the sets have many spoligotypes in common and are roughly of the same size [55]. Figure 4.5 graphically explains how the best match values were calculated. For each cluster in each model, we calculate the best match with respect to all other models. After that, we average these best match values over all of the clusters in a particular model with respect to the rest of the models, thus obtaining the stability of the model. To calculate the stability of a cluster we average the cluster's best match values over all of the generated models.

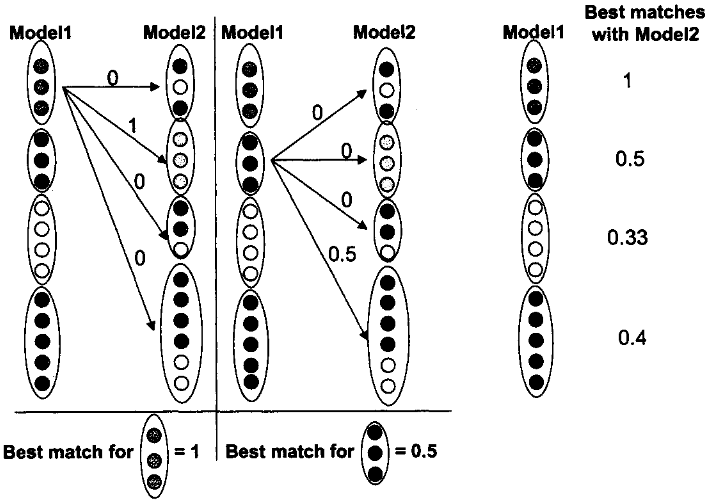


Figure 4.5: Explanation of calculation of best math values

The stabilities of the families produced by a model initialized with the SpolDB3-based prototypes were assessed by comparing the 36 identified families with the families identified by the 100 36-order randomly initialized models, which is shown in Fig. 4.6. These models were initialized as schematically shown in Fig. 4.3.

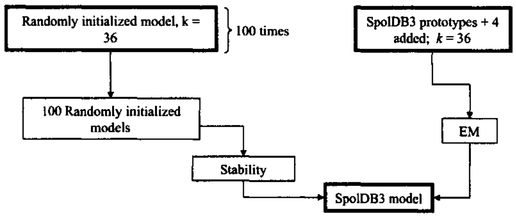

Randomly initialized model, k = 36
100 times
SpolDB3 prototypes + 4 added; k = 36
100 Randomly initialized models
EM
Stability
SpolDB3 model

Figure 4.6: Schema of the analysis of the clusters identified by the SpolDB3-initialized model

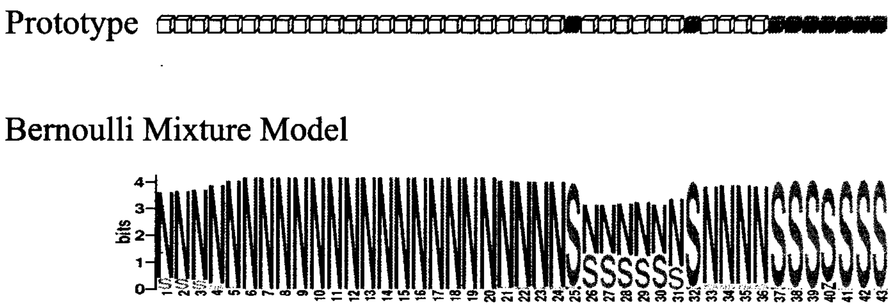
4.5 Results

4.5.1 Families identified using SpolDB3-based model

At first, we applied the Bernoulli mixture model initialized with the 32 SpolDB3-derived prototypes to the data set including 845 unique and 535 shared patterns. Some of the resulting families did not reflect the current view on spoligotypes' evolution, i.e. their members had spacers in the positions where the prototype for this family did not have spacers and therefore could not be an ancestor of these strains.

Figure 4.7 shows a SpolDB3-based prototype and a sequence logo for one such family where the spacers were "acquired" by the children of the family's prototype: some of the spoligotypes in the family had spacers at positions 1-3 and 26-31, at the same time as their parent spoligotype did not. Sequence logos are a common graphical representation of an amino acid or nucleic acid multiple sequence alignment [20, 111]. They illustrate the location and degree of sequence conservation in the set of aligned sequences. Sequence logo analysis was previously carried out on spoligotyping data and was found to be useful in representing selected groups of spoligotypes and examining phylogenetic relationships of the MTC strains [26].

Some of the spoligotype patterns in the analyzed database did not correspond to any of the SpolDB3-defined prototypes; therefore, these previously empirically defined prototypes were supplemented with four additional prototypes. The orphan (unique) spoligotype patterns were excluded from further analysis, since their



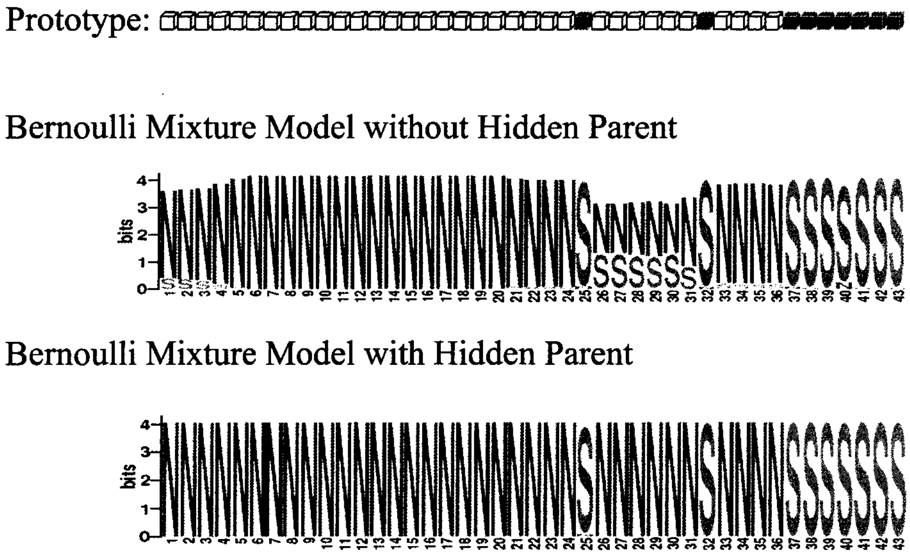
Prototype

Bernoulli Mixture Model

Figure 4.7: Prototype and sequence logo for the *M. tuberculosis* Haarlem2 family identified from the whole data set using Bernoulli mixture model; S indicates spacer, N indicates absence of spacer

validity may be questionable.

The EM algorithm was modified by introducing the Hidden Parent into the Bernoulli mixture model, which resulted, when EM was started with the same initial parameters, in families that were more consistent with the biologically relevant definition of a spoligotype family being a collection of children of a Hidden Parent (see Fig. 4.8). We adopted the Hidden Parent model for all of our further experiments presented here.



Prototype:

Bernoulli Mixture Model without Hidden Parent

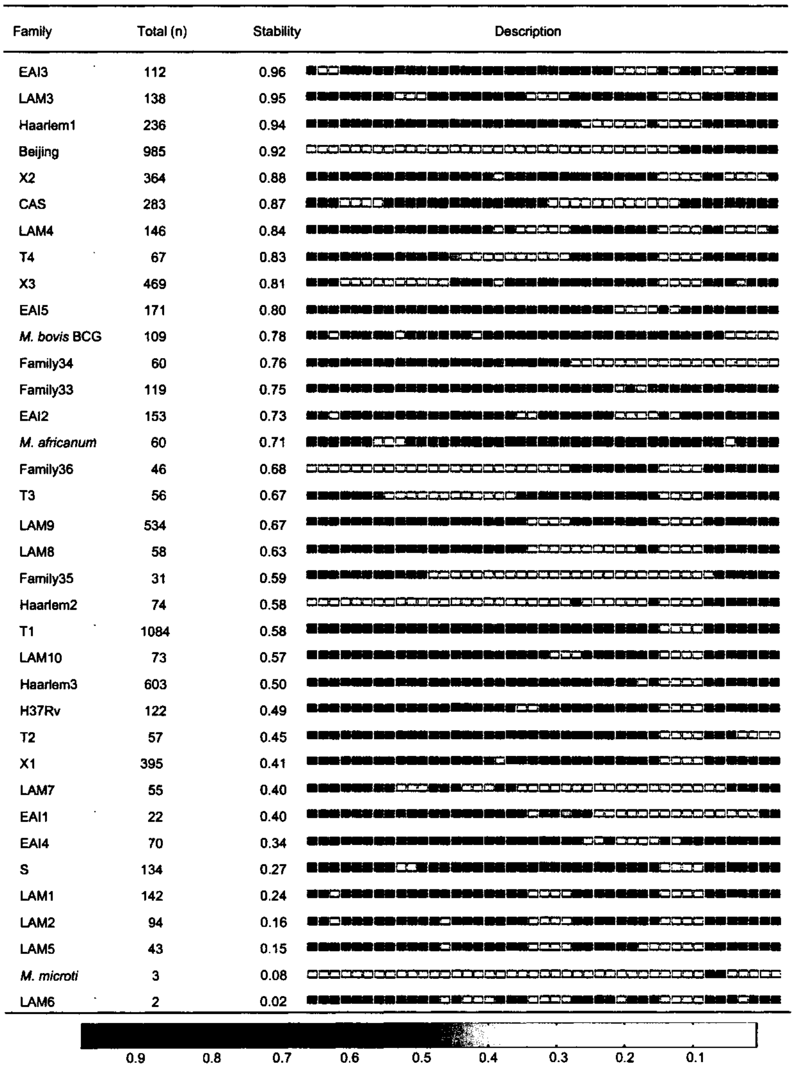
Bernoulli Mixture Model with Hidden Parent

Figure 4.8: Difference in the content of *M. tuberculosis* Haarlem2 family identified with and without Hidden Parent; S indicates spacer, N indicates absence of spacer

Figure 4.9 describes the families identified using the SpolDB3-defined prototypes plus four prototypes added in this study for initialization. The first column contains the names for spoligotype families as defined in [35], as well as families 33 – 36, resulting from the 4 additional prototypes. The second column shows the total number of isolates, each corresponding to a TB patient, in the family. The third column contains the stability value for each family, and the fourth is a schematic representation of probabilities of the 43 spacers in spoligotypes within the family, i.e. the Hidden Parent of the family. The colorbar underneath the table shows the gradation of colors corresponding to probabilities of spacers.

Each among the 533 isolates was assigned to one or another among the 36 possible families with probability greater than 0.5. Spoligotypes describing two remaining shared types, with octal codes designations [22] 776377777720771 and 776377777420771 each belonged with approximately equal (0.44) probability to families Haarlem3 and S, and with probability 0.12 to family T1. It is worth noting that the *M. bovis*-BCG family contains mostly isolates of *M. bovis* strains. The canonical *M. bovis*-BCG spoligotype has octal code 676773777777600, and the New York State database contains 41 isolates with this spoligotype out of the total of 109 isolates in the *M. bovis*-BCG family. We keep the *M. bovis*-BCG name for this family to be consistent with the SpolDB3 notation.

We also report the stability values for each of the 36 families relative to the 100 randomly initialized models each having 36 components. When compared to these 100 clustering solutions, 23 families of the SpolDB3 clustering had stability values higher than 0.5, five families, Haarlem3, H37Rv, T2, X1 and LAM7, had stabilities between 0.4 and 0.5, and the rest (EAI1, EAI4, S, LAM1, LAM2, LAM5, LAM6 and *M. microti*) were not stable, with stability values below 0.4. Interestingly, the stabilities of defined in this work families 33 and 34 (temporarily numbered so for convenience purposes) were quite high. Family33 included a shared type of size 22 wherein only two spacers, 33 and 34, were absent. This type was recently described as belonging to a clade MANU [117]. Family 33 gathers spoligotypes with most of the spacers present; this spoligotyping pattern is the closest to that of a putative common ancestor.



Family	Total (n)	Stability	Description
EAI3	112	0.96	
LAM3	138	0.95	
Haarlem1	236	0.94	
Beijing	985	0.92	
X2	364	0.88	
CAS	283	0.87	
LAM4	146	0.84	
T4	67	0.83	
X3	469	0.81	
EAI5	171	0.80	
M. bovis BCG	109	0.78	
Family34	60	0.76	
Family33	119	0.75	
EAI2	153	0.73	
M. africanum	60	0.71	
Family36	46	0.68	
T3	56	0.67	
LAM9	534	0.67	
LAM8	58	0.63	
Family35	31	0.59	
Haarlem2	74	0.58	
T1	1084	0.58	
LAM10	73	0.57	
Haarlem3	603	0.50	
H37Rv	122	0.49	
T2	57	0.45	
X1	395	0.41	
LAM7	55	0.40	
EAI1	22	0.40	
EAI4	70	0.34	
S	134	0.27	
LAM1	142	0.24	
LAM2	94	0.16	
LAM5	43	0.15	
M. microti	3	0.08	
LAM6	2	0.02	

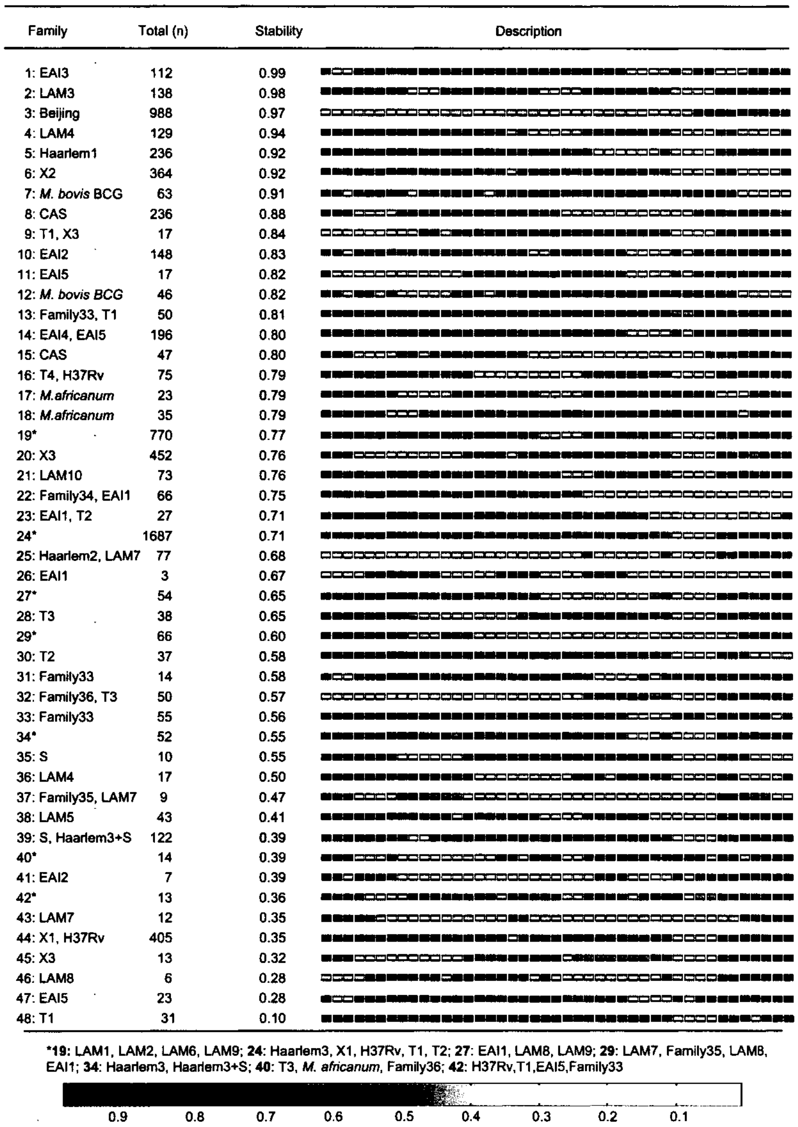
Figure 4.9: Summary of the results obtained using the SpolDB3-derived prototypes for model initialization. Probability of a spacer in Hidden Parent is represented by colored box where gradation of colors corresponds to probabilities of the spacer's presence: white indicates 0 and black indicates 1

Results of a recent work on identification of MTC isolates by chromosomal deletion analysis confirmed the identity of 12 *M. africanum* strains each having distinct spoligotype [99],[Parsons, personal communication]. Seven of these 12 spoligotypes were in our training database and six of the seven were correctly identified using the SpolDB3-based model as *M. africanum.* The seventh was placed in family 35 because it has very unusual for the *M. africanum* family spoligotyping pattern, with spacers 10-37 absent. Five strains were absent from our database and therefore could not be used to train the SpolDB3-based model. Of the five, when submitted to SPOTCLUST, three were correctly identified as *M. africanum*, and two were assigned to families T2 and 33, again because their spoligotypes were very different from the SpolDB3-derived definition of family *M. africanum.* Under the assumption of the SPOTCLUST and SpolDB3 expert rules, *M. africanum* strains belong to more than one spoligotyping family.

4.5.2 Families identified using randomly initialized model (RIM)

For the RIM, we needed to determine the number of mixture components, or the model order. Figure 4.11 demonstrates the results of the application of MCCV to our spoligotyping data. The average test log-likelihood over 100 different cross-validation partitions are plotted against the model orders. We have chosen 48 to be the optimal model order, since this point corresponds to a peak in the average test log-likelihoods. Moreover, after this point the curve levels off. This indicates that a further increase in the number of parameters will not significantly improve the log-likelihood [131, 139].

The 48 families identified by the RIM, sorted by their stability values, are summarized in Fig. 4.10. We observed that on average the total log-likelihood of a model increases with the stability of the model. Therefore, out of the 100 randomly initialized 48-order mixture models, we chose as a final solution the one that, when fitted to the data, allowed EM to converge within 300 iterations and achieve the highest total log-likelihood. Simultaneously, this model was characterized by the highest stability. The first column in this figure contains the spoligotype family labels given to them by the model initialized with the SpolDB3-derived prototypes



Family	Total (n)	Stability	Description
1: EAI3	112	0.99	
2: LAM3	138	0.98	
3: Beijing	988	0.97	
4: LAM4	129	0.94	
5: Haarlem1	236	0.92	
6: X2	364	0.92	
7: *M. bovis* BCG	63	0.91	
8: CAS	236	0.88	
9: T1, X3	17	0.84	
10: EAI2	148	0.83	
11: EAI5	17	0.82	
12: *M. bovis BCG*	46	0.82	
13: Family33, T1	50	0.81	
14: EAI4, EAI5	196	0.80	
15: CAS	47	0.80	
16: T4, H37Rv	75	0.79	
17: *M.africanum*	23	0.79	
18: *M.africanum*	35	0.79	
19*	770	0.77	
20: X3	452	0.76	
21: LAM10	73	0.76	
22: Family34, EAI1	66	0.75	
23: EAI1, T2	27	0.71	
24*	1687	0.71	
25: Haarlem2, LAM7	77	0.68	
26: EAI1	3	0.67	
27*	54	0.65	
28: T3	38	0.65	
29*	66	0.60	
30: T2	37	0.58	
31: Family33	14	0.58	
32: Family36, T3	50	0.57	
33: Family33	55	0.56	
34*	52	0.55	
35: S	10	0.55	
36: LAM4	17	0.50	
37: Family35, LAM7	9	0.47	
38: LAM5	43	0.41	
39: S, Haarlem3+S	122	0.39	
40*	14	0.39	
41: EAI2	7	0.39	
42*	13	0.36	
43: LAM7	12	0.35	
44: X1, H37Rv	405	0.35	
45: X3	13	0.32	
46: LAM8	6	0.28	
47: EAI5	23	0.28	
48: T1	31	0.10	

***19**: LAM1, LAM2, LAM6, LAM9; **24**: Haarlem3, X1, H37Rv, T1, T2; **27**: EAI1, LAM8, LAM9; **29**: LAM7, Family35, LAM8, EAI1; **34**: Haarlem3, Haarlem3+S; **40**: T3, *M. africanum*, Family36; **42**: H37Rv,T1,EAI5,Family33

Figure 4.10: Summary of the results obtained using RIM. Probability of a spacer in Hidden Parent is represented by colored box where gradation of colors corresponds to probabilities of the spacer's presence: white indicates 0 and black indicates 1

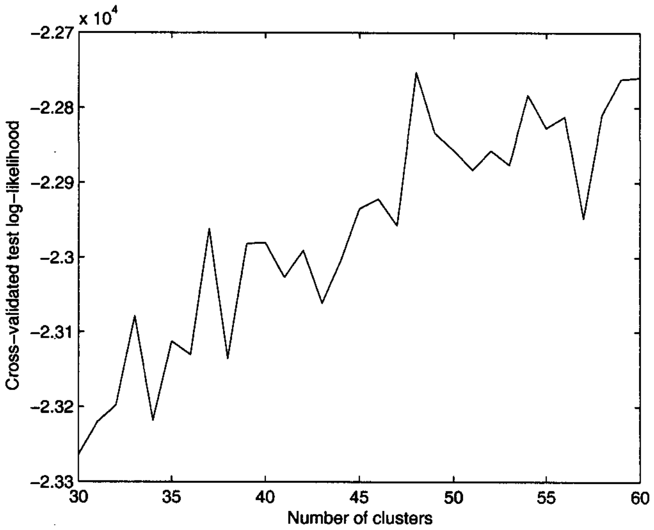


Figure 4.11: MCCV results: Cross-validated test log-likelihoods versus number of clusters for $k = 30, \cdots, 60$

and the prototypes for families 33-36. The 48 families are numbered for convenience; we will refer to them as the N families.

Of the 48 families, 35 have average stability values greater than 0.5. Another 13, while they are not stable, with stabilities between 0.1 and 0.5, nevertheless occur in the same content in several other high-stability and high-log-likelihood models.

Most of the highly stable families identified by the RIM corresponded to the highest-stability families produced from the SpolDB3 prototypes: EAI3, LAM3, Beijing (included *Mycobacterium microti*), Haarleml, LAM4, X2, X3, EAI2 and LAM10.

Some of the families split into two: the *M. africanum* spoligotypes formed two distinct families corresponded to the SpolDB3 prototypes for classes Afril and Afri2–3, respectively (see families N18 and N17 in Fig. 4.10). The CAS spoligotypes also split into two stable families N8 and N15 matching the SpolB3 prototypes for CAS1 and CAS2. Spoligotypes previously placed in families 33 and 34 formed two stable families N13 and N22, respectively, each having the content almost identical to that of their corresponding counterparts resulting from the use of the SpolDB3-based seeds.

Several SpolDB3-derived families merged. The largest of them was the 0.71-stability value family comprising spoligotypes labeled as belonging to families T1, Haarlem3, X1, and H37Rv (family N24 in Fig. 4.10). Similarly, members of several LAM families aggregated into one family with stability value 0.77 (family N19). Spoligotypes labeled as belonging to the EAI4 family, and some EAI5 spoligotypes, also formed one stable family. Family S was reproduced nicely by RIM, but with a low stability. Haarlem3-labeled spoligotypes formed a separate family only if the spacers 29-31 and 33-36 were absent simultaneously. Several medium- and low-stability clusters, which we could only tentatively call families, contained a few shared types or only a single one.

4.6 Discussion

4.6.1 SpolDB3-initialized model

The initial mixture components derived from the international spoligotyping database SpolDB3 were based on the visual inspection of the spoligotype profiles by a human expert [113]; therefore, they fitted the structure of the data well. The algorithm was forced to identify manually defined MTC strain families. To evaluate the bias in the identification of these families, we randomly initialized 100 36-component models and compared the resulting families with each of the SpolDB3-based families. Out of the 36 SpolDB3-based families, 15 had stability greater than 0.7 (see Fig. 4.9).

A SpolDB3-defined family was reported as stable if it was reproduced in the identical content by the majority of the 100 36-component RIMs. In other words, the stability value for a family represents the frequency with which our algorithm finds this family given that it has initially no knowledge about the existent families except their total number, 36. The stability is of importance to us because we try to minimize the human input into the spoligotyping data analysis and this value helps us to assess our efforts. High stability value of a SpolDB3-based family means that the family is well reproducible by our algorithm; low manifests that the algorithm is unable to consistently identify the family. Low stability indicates that either the family is not well defined or the model needs to be improved. Both of these factors may influence the results as well. We observed that SPOTCLUST does quite a good job of finding expert-defined families, not without some exceptions, however, which we will discuss shortly.

Use of the Bernoulli mixture model with Hidden Parent as opposed to the model without Hidden Parent resulted in biologically more correct families, because the child spoligotype was allowed to have some spacers lost but not gained, reflecting currently widely accepted hypothesis on evolution of the DR locus. In fact, with the exception of a few shared types, the spoligotypes in the 36 families were legitimate children of their expert-defined prototypes. In family LAM8, three spoligotypes had spacer 21 revealed. However, the rest of their pattern matched that of their parent. Because there was no other, more suitable, parent for these spoligotypes, they were forced to join this family.

In general, SPOTCLUST includes each spoligotype in the closest family, complying with the hypothesis of losing spacers by the DR locus as it evolves. Upon including new spoligotypes in a family, SPOTCLUST changes the parameters of the family's Hidden Parent to accommodate for the existence of the new members, thus predicting a new, legitimate to all of the family members, Hidden Parent. The Hidden Parent is not a spoligotype, but a probabilistic entity. Most of the SpolDB3-based families contain a spoligotype that could be considered ancestral to the rest of the members. For example, in family LAM10, the pattern with octal code 777777743760771 is parental to other spoligotypes that only miss spacers compared to their parent. Some families have a hypothetical parent only.

SPOTCLUST, with random initialization, did not identify several of the SpolDB3-based families as distinct ones. Some SpolDB3-based families had low stabilities. Since some of the SpolDB3-derived prototypes were children or parents of other prototypes, the randomly initialized EM did not always distinguish these families. Families such as X1, H37Rv, S, EAI2, T2 and Haarlem3, had low, less than 0.5, stability values relative to the randomly initialized 36-order models. Even when the EM algorithm was randomly initialized with a 100-component model, these families merged into a single family whose Hidden Parent would be a spoligotype exactly matching the prototype for family $T1$ (data not shown). In general, the higher the number of the components in the model, the harder EM tries to divide large families into smaller ones; therefore, the fact that the families were not separated in a 100-component model serves as an indication that according to our method these families comprised one big family.

Most of the spoligotypes in the T families had many spacers missing respective to their initial parents, the SpolDB3-derived prototypes. Some of the spoligotypes included in family $T1$ matched the prototype for family T2, a not surprising result, since the prototype for T2 is the child of that for T1 (see Fig. 4.1). The T family is currently considered "ill-defined" [28]. We were unable to create models that discriminated well among its members; therefore, our conclusions concurred with referring to this family as poorly defined.

Our results suggested that some of the SpolDB3-based prototypes were redundant and should be restated for use in the context of our approach, perhaps in a hierarchical fashion. When 36-order models were randomly initialized to identify strain families, spoligotypes from all LAM families, except LAM3 and LAM4, merged into one family composed of children of LAM9 (see Fig. 4.1 for prototypes for these families). The LAM3 family was very stable, probably because there were enough almost identical shared types to form this distinct set. The *M. microti* family contained only one shared type wherein all spacers, except 37 and 38, were absent. This family had a very low stability, since in the randomly initialized 36-order model the *M. microti* shared type was always included in the Beijing family. Spoligotypes that have been allocated to families EAI1 and EAI2 by the SpolDB3-derived model were mixed with spoligotypes from other families (33, 34, and sometimes LAM8) when the model was initialized randomly; therefore, the stabilities of EAI1 and EAI2 were low.

We can conclude the discussion with the suggestion that further analysis is needed to determine what exactly constitutes a spoligotyping family. Our results confirm most of the families previously distinguished within the SpolDB3 database and indicate which families require special attention.

4.6.2 RIM

Finding the "optimal" number of different groups in the data, without any prior information, presents a rather challenging, sometimes unrealistic, problem. The solution is highly dependent on the algorithm used, the model initialization, the data characteristics, and the definition of the "optimality" itself [118]. Using the MCCV approach to determine the number of distinct families in our spoligotyping data, we have concluded that 48 represented a reasonably good number of components in the model (Fig. 4.11). The highest total log-likelihood and total stability were criteria for validation of our algorithm. We consider the final model to be the one with parameters best fitting our data. It should be noted, however, that because our method employs probabilistic models, the correct number of mixture components and their parameters do not exist as single numbers, but instead each vary within a certain range. We do not claim that there are exactly 48 spoligotyping families; we show only that in the context of our model definition this number reflects well the structure of the data. The solution that we report here should be considered as probabilistically good, given our choice of method.

The majority of the stable families identified by the SpolDB3-based model were also identified by the RIM. Some of the SpolDB3-defined families merged into one family (see, for example, families N19, N24, N27, N40, and N42 identified by the RIM) and thus could be considered to have potentially independently evolved from the same ancestral strain. This is the same conclusion that we had made upon analysis of the SpolDB3-based families and their stability. Our conclusions again concurred with the previous reporting of T family as poorly defined [28]. Taken together, this indicates that, if we are to preserve the SpolDB3 recognition rules, the Hidden Parent model may need to be refined, possibly by introduction of a hierarchy concept into the model, or by separately identifying subfamilies within certain big families. Some novel families, such as families N9, N13, N14, N16, N19, N22, N23 and N24, each characterized by a newly defined Hidden Parent, were stable. Appearance of small families may be due to the current lack of genotyped MTC strains, even though most of the samples in our collection were from NYC, whose TB patient population is one of the most diverse in the United States.

We can conclude that the RIM distinguishes the major spoligotyping families well, while suggesting some new families whose validity needs to be further examined.

The fact that some of the families identified by SPOTCLUST are not stable, suggests room for further refinement of the model. One possible improvement of our algorithm is the incorporation of interdependencies of spacers. Results of previous work suggested that the deletion of contiguous DVR sequences does not occur sequentially, but rather by a single loss of several adjacent DVRs [3, 154]. This severely complicates the use of spoligotypes for the derivation of MTC phylogeny [154]. There is evidence that a canonical Beijing spoligotype appeared as a result of one event that was a simultaneous loss of 34 contiguous DVRs from an ancestral spoligotype initially having all of the 43 DVRs present (N. Kurepina, personal communication). Also, some spacers (for example, 33-36) are simultaneously absent in most spoligotypes, a feature which may indicate their interaction. Another complication arises if some spacers were lost independently in distinct families resulting in convergent spoligotypes. Moreover, certain spacers may be present but undetected by spoligotyping due to particular IS*6110* insertions [91]. We should therefore consider these factors when inferring the parent-child relationships of spoligotypes.

CHAPTER 5
Analysis of demographic data on patient isolates in MTC strain families

This chapter is dedicated to the assessment of possible advantages associated with identifying of the MTC strain families based on spoligotyping data. We examined the available patient data accompanying each data point within the MTC strain families.

5.1 Methodology and available data

Our ultimate goal of designing a decision-making tool for TB control purposes required fusing information from TB strain genotyping and demographic patient data. Here we showed how analyzing patient data by the identified spoligotype families can yield valuable insights into underlying disease trends. We analyzed the NYC spoligotyping database that comprised isolates collected from January 1, 2001 to July 1, 2004. It included 220 shared types for 2297 isolates and 389 unique spoligopatterns. Each of the isolates was annotated with patient's age, gender and country of birth; for each foreign-born patient, the date of his/her entry to the United States (US) was available. The NYC database contained information for patients from 112 countries; we grouped the countries other than the US into eight geographic regions as follows. Central America: Belize (n = 3), Guatemala (n = 8), El Salvador (n = 13), Honduras (n = 24), Mexico (n = 123), Nicaragua (n = 2), Panama (includes Canal Zone) (n = 6); South America: Argentina (n = 5), Bolivia (n = 4), Brazil (n = 11), Chile (n = 1), Columbia (n = 30), Ecuador (n = 184), Guyana (n = 48), Peru (n = 53), Uruguay (n = 1), Venezuela (n = 1); the Caribbean: Aruba (n = 1), Bahamas (n = 1), Barbados (n = 1), Cayman Islands (n = 1), Cuba (n = 13), Dominica (n = 1), Dominican Republic (n = 126), Grenada (n = 2), Jamaica (n = 15), Puerto Rico (n = 75), St. Kitts and Nevis (n = 1), St. Lucia (n = 1), St. Vincent and the Grenadines (n = 2), Trinidad and Tobago (n = 21), U.S. Virgin Islands (n = 3), Haiti (n = 112); Europe: Albania (n = 5), Austria (n = 3), Belarus (n = 1), Estonia (n = 1), Finland (n = 1), Germany (n = 2), Greece (n = 1), Hungary (n = 1), Ireland (n = 2), Italy (n = 4), Lithuania (n = 1), Macedonia (n = 2), Poland (n = 9), Portugal (n = 3), Romania (n = 7), Russia (n = 19), Spain (n = 3), Turkey (n = 6), Ukraine (n = 11), Yugoslavia (n = 8); Africa: Angola (n = 2), Burkina (n = 1), Cameroon (n = 4), Central African Republic (n = 1), Chad (n = 1), Egypt (n = 5), Ethiopia (n = 7), Gambia (n = 11), Ghana (n = 15), Guinea (n = 22, Ivory Coast (n = 13), Kenya (n = 1), Liberia (n = 10), Malawi (n = 1), Mali (n = 19), Mauritania (n = 2), Morocco (n = 3), Niger (n = 3), Nigeria (n = 18), Senegal (n = 15), Sierra Leone (n = 6), Somalia (n = 1), South Africa (n = 5), Sudan (n = 1), Tanzania (n = 2), Togo (n = 6), Tunisia (n = 1), Zambia (n = 8), Zimbabwe (n = 2); Central Asia and Middle East: Afghanistan (n = 3), Armenia (n = 1), Bangladesh (n = 31), Bhutan (n = 1), Georgia (n = 2), India (n = 117), Kazakhstan (n = 1), Nepal (n = 37), Pakistan (n = 53), Saudi Arabia (n = 2), Sri Lanka (n = 1), Turkmenistan (n = 1), Yemen (n = 7); Far East: Cambodia (n = 5), China (n = 259), Hong Kong (n = 21), Indonesia (n = 15), Japan (n = 5), Macau (n = 2), Malaysia (n = 3), Mongolia (n = 1), Myanmar (n = 12), North Korea (n = 3), Philippines (n = 65), South Korea (n = 70), Taiwan (n = 6), Thailand (n = 6), Vietnam (n = 26). The database also contained three Canadian-born patients that were considered separately. A total of 758 isolates were obtained from US-born patients.

We applied both the SpolDB3-based and randomly initialized models trained on the New York State database to the smaller NYC database, which was, with the exception of 247 orphans and seven shared types, a subset of the former. Orphan spoligotypes were excluded from the analysis when the models were trained on the New York State database. However, to test the models, it was appropriate to include orphans (unique spoligopatterns), because one of our goals was to be able to make inferences about orphan spoligotypes given knowledge acquired from studying shared types. Out of 389 orphans in the NYC database, 126 were present at least twice in the New York State database, thus being shared types in the latter. The model trained on the larger database was fixed, i.e. its parameters were considered final. Each spoligopattern in the NYC database was then assigned to its most probable family. We analyzed patient age at TB diagnosis, gender, geographic origin, and the time foreign-born patients had spent in the US before the advent of the infection, with respect to the identified families.

5.2 Results and Discussion

We present the results of the patient data analysis for the MTC strain isolates in the families identified using the SpolDB3-based model. We limited our discussion to the families identified using this model, since the prototypes for these families were previously expert-defined based on the global international spoligotyping database [35]. Moreover, the major families have been described by different research groups studying MTC isolates obtained from TB patients from different countries [1, 14, 81, 113, 117].

The isolates comprising the NYC database were grouped into the 36 families. All except one of the resulting families contained at least one isolate: none of the isolates was identified as belonging to the *M. microti* family. Family LAM6 contained only one isolate. The 15 stable families, EAI3, LAM3, Haarlem1, Beijing, X2, CAS, LAM4, T4, X3, EAI5, *M. bovis*-BCG, 33, 34, EAI2, and *M. africanum*, characterized by stability values greater than 0.7, represented the most robust cases, since the RIM identified them in the same content as the SpolDB3-based model. Nevertheless, we discuss all of the 35 families. The low stability value indicates that the family would be difficult to identify by our method given little or no prior knowledge about the family; however, the family may still represent a natural grouping.

Analysis of the data on US-born and foreign-born individuals showed that the dynamics of transmission of MTC isolates within these two groups of patients varied with the identified families. In the NYC database, the number of TB cases among foreign-born persons prevailed over that of the US-born infected persons (72% versus 28%, respectively).

Figure 5.1 shows the variation in the number of US- and foreign-born patients in different MTC strain families. In the majority of the families (27/35), isolates from foreign-born patients clearly prevailed. The histograms for four families, X2, X3, LAM4 and LAM5, demonstrate the predominance of US-born patients; in LAM4 this prevalence is particularly strong. Small families T4, T3 and EAI1, as well as a medium size family LAM8, comprise roughly equal number of strains isolated from patients born inside and outside the US.

The groups comprising US- and foreign-born patients can be divided into subgroups of clustered and unique cases. In an epidemiological language, a cluster is defined as a set of two or more isolates recovered from different patients and possessing identical genotypes. In other words, each cluster is a shared type observed in the NYC database. We were interested in distinguishing clustered and unique cases for the following two reasons. First, it is widely assumed that clustered cases are more likely to be directly or indirectly involved in the same chain of TB transmission, while unique cases are more likely to result from the reactivation of latent infection [10]. The second reason was that when an isolate with a unique genotype is encountered, it is difficult to make a plausible suggestion on its origin. When the unique isolate belongs to a particular family, we can draw inferences about this case based on the information about other strains in the family.

Figure 5.1 depicts the total number of cases that belong to one or another shared type as opposed to the isolates that have no match in our database. We will call the latter isolates non-clustered or unique. However, one should remember that they are unique given the NYC database only; other databases may contain these isolates as well. The dissimilarities in the number of clustered and non-clustered cases in different families are apparent. In the overwhelming majority of the families, the non-US-born clustered cases predominate. The *M. africanum* and EAI5 families showed unusually large percentages of unique spoligotypes from strains infecting non-US-born patients. Taken together with the observation, which is discussed below, that these patients immigrated from their countries of origin relatively recently, this fact suggests that the TB transmission occurred primarily outside of the US. About a third of the families (LAM3, Haarlem1, X2, T4, Family33, LAM9, LAM8, Haarlem3, X1 and S) contain nonnegligible number of isolates with unique spoligotypes obtained from US-born persons. These cases should be given particular attention since they most probably indicate the recent transmission of TB. Besides, TB control measures are more effective when targeted toward US-born patients

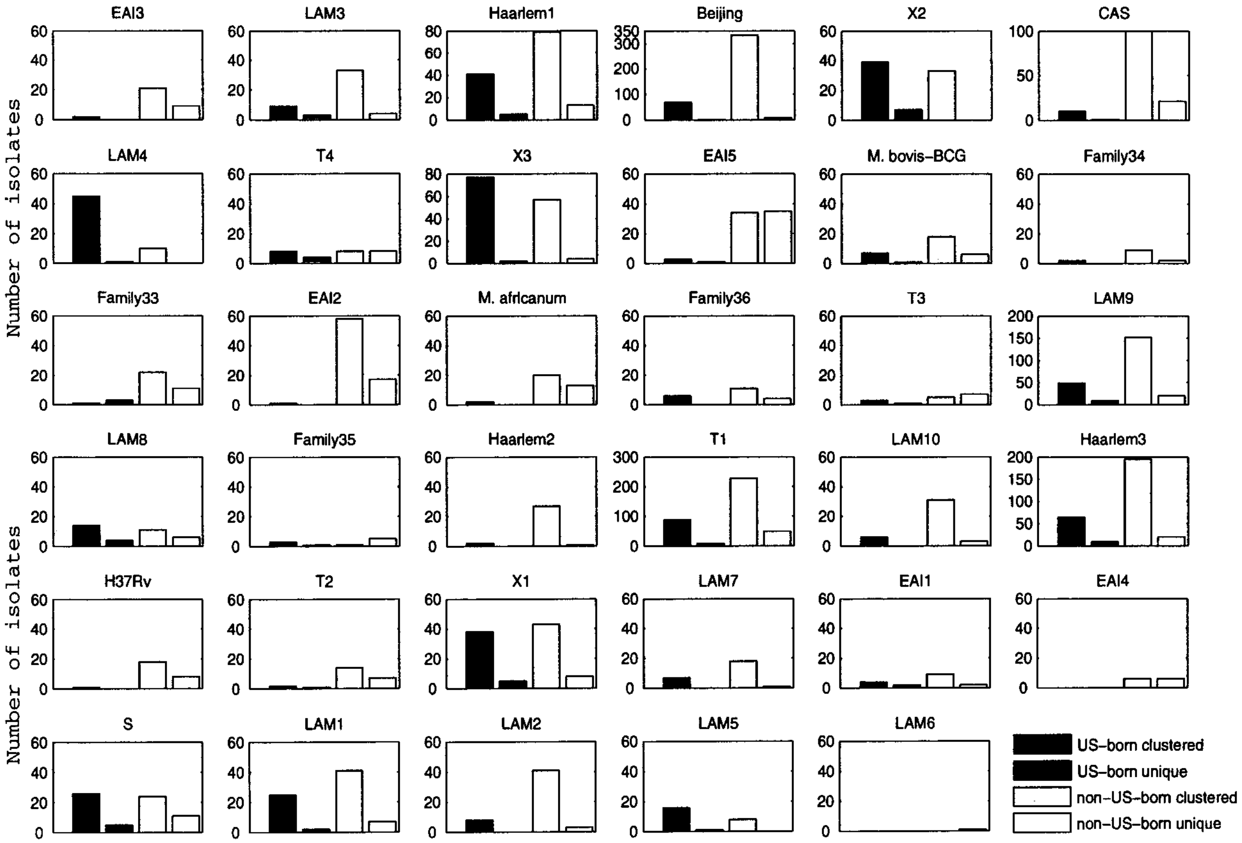


Figure 5.1: Number of clustered and non-clustered isolates in groups of US- and foreign-born TB patients, by MTC strain families identified in NYC database

[45, 79]; therefore, an outbreak can be stopped at an early stage. The described trends can be further investigated by examining the age and immigration date for each foreign-born patient.

The majority of non-US-born patients carrying MTC strains that belong to large strain families originated from particular geographic regions. Figure 5.2 shows geographical origin of patients in the 35 MTC strain families. EAI3 and CAS contain strains isolated mostly from patients that came from the Middle East and Central Asia. Isolates obtained from patients born in these regions are also abundant in EAI5 and family 33. Members of the Haarlem2 and LAM2 strain families were isolated mainly from the patients originated in the Caribbean. Patients infected with isolates included in family LAM4 were predominantly born in the US. More than half isolates in the *M. bovis*-BCG family were obtained from the Mexican-born patients. The majority of the Beijing and EAI2 isolates were collected from patients born in Far East countries. Isolates from Africa-born patients dominated, as the name implies, in the *M. africanum* family. The majority of the isolates from family LAM10 are also obtained from African-born patients. Several big families, such as Haarlem1, LAM9, T1 and Haarlem3, encompass MTC strains isolated from patients that represent all of the eight geographic regions considered . The isolates from three Canadian-born patients are a little difficult to see on Fig. 5.2: each of them belongs to a different family: X2, LAM9, and Haarlem3.

Examination of the duration of time spent in the US by foreign-born patients before they were diagnosed with TB revealed that most of the immigrants in the identified families had been in the country for less than 20 years by the time they developed the active disease. This indicated that, most probably, the majority of TB cases among the foreign-born persons were due to the reactivation of the latent infection, which was previously shown for NYC [45] and Massachusetts [116]. There is also a possibility that the immigrants acquire new infections or reinfections, either through transmission within the US, which may be associated with their residing in communities populated by other immigrants, or through frequent visits to their country of origin. This scenario was suggested to explain high incidence rates of TB among the immigrants in the Netherlands a decade after their immigration date

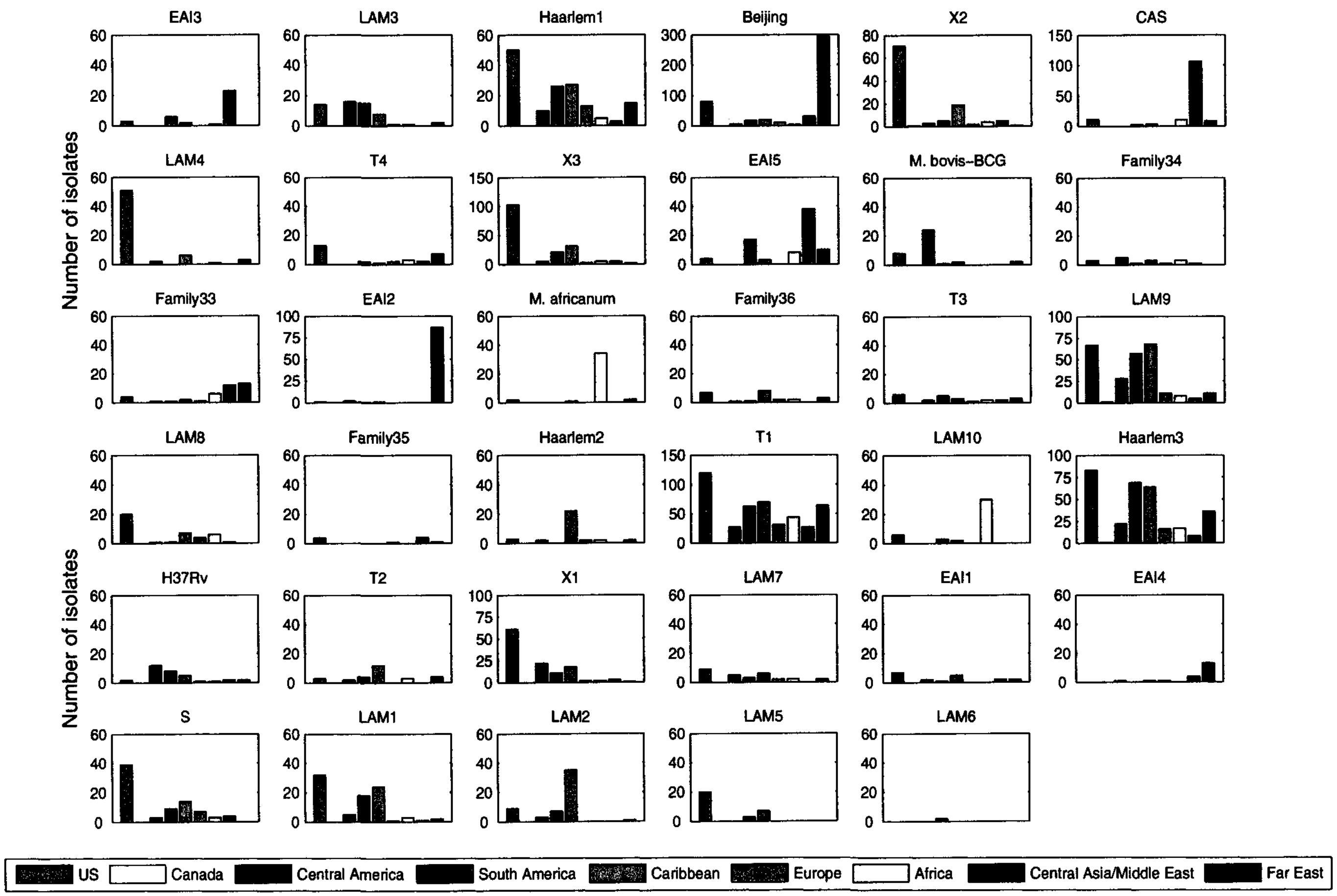


Figure 5.2: Analysis of geographical origin of patients by MTC strain families identified in NYC database

[151]. Figure 5.3 shows the age and time spent in the US at the advent of TB by the immigrants from different geographic regions. The green dots appearing on the diagonal show the age of the US-born persons carrying MTC strains that belong to the family. This figure illustrates the overall distribution of MTC strain families by geographic regions. Analysis of families EAI3, *M. africanum*, CAS, EAI2, LAM10 and EAI4 showed that most isolates within them were obtained from non-US-born patients that have been in the country for less than 20 years and each came from a particular geographic region. This suggested that these patients contracted TB before coming to the US [45]. In contrast, examination of isolates from Beijing, Haarlem1, LAM9, T1, and Haarlem1 families revealed that many (but not the majority) of them were recovered from foreign-born patients of various ages that have been in the country for more than 20 years. These infections may have been acquired in the US; alternatively, strains from these families possess higher ability to host adaptation [43, 54]. We can observe in Fig. 5.3 that in NYC, on average, foreign-born TB patients were younger than US-born. The patients infected with *M. bovis*-BCG strains were unusually young, which is elaborated in more details below.

Ordinarily, it is easier to elucidate the dynamics of TB transmission among US-born patients than among foreign-born ones [45]; therefore, it is more informative to assess the age distribution among US-born patients. Figure 5.4 allows us to examine the age of US-born persons infected by the MTC strains. This figure also shows the number of US-born patients within each family. The median and average age of the studied US-born population were both 45. The age distribution by families noticeably varied. Some families, such as CAS and *M. bovis*-BCG, contained isolates from unusually young US-born individuals, which suggested further investigation of these groups. Families *M. africanum*, 34, T2 and Haarlem2 each contained a small number of isolates from patients aged over 60 years. Family *M. bovis*-BCG presented a very interesting case. Our results demonstrated that: a) the majority of TB patients in this family are from Mexico; and b) US-born patients in this group are very young. It turns out that this family contained isolates from persons infected as a result of an outbreak occurred mostly among Mexico-born NYC residents and US-

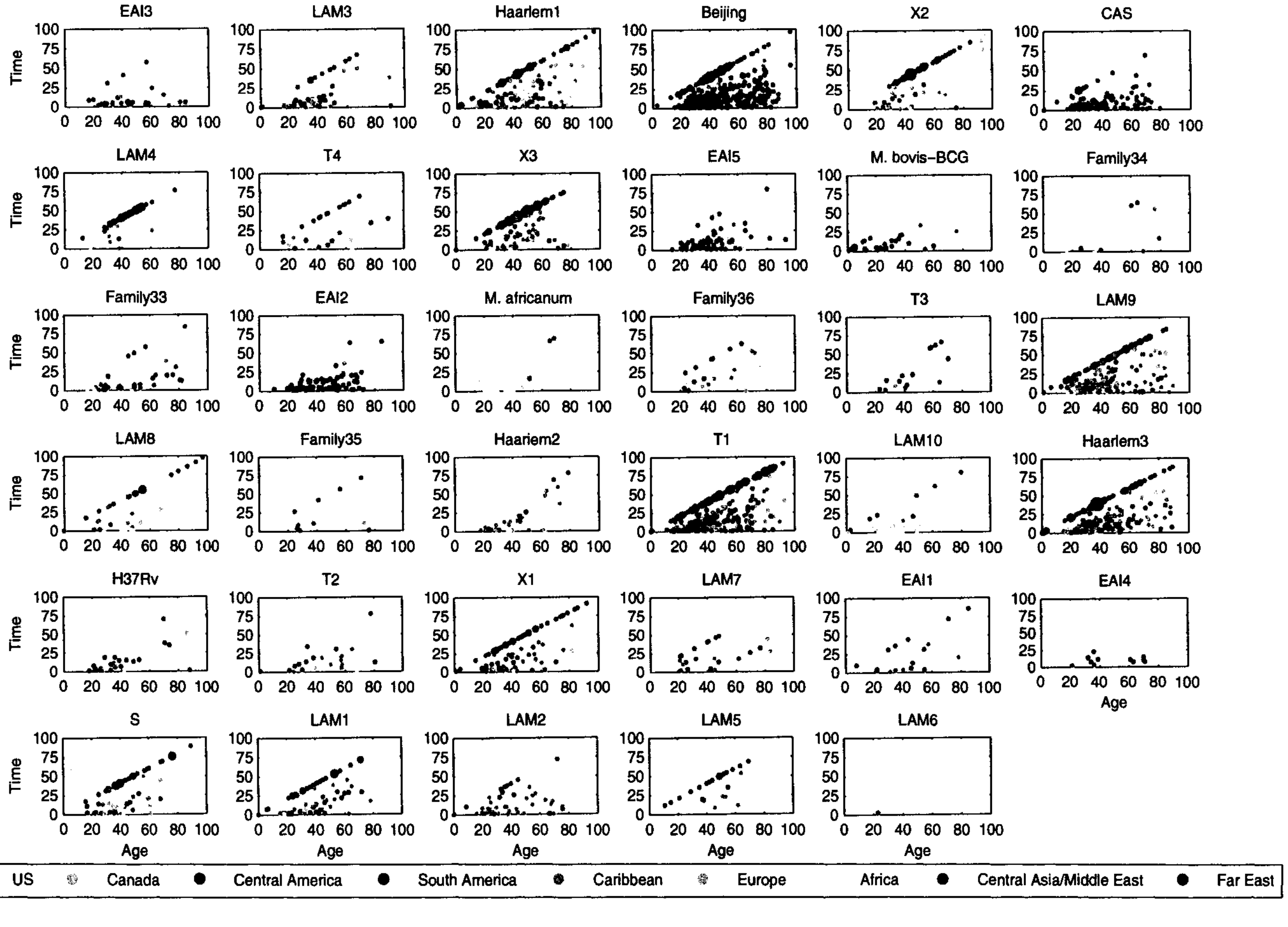


Figure 5.3: Time in the US versus age at TB diagnosis for foreign-born patients by their geographic origin. The color of the dot representing a patient corresponds to the geographic region where this person was born. The age of the US-born patients is plotted on the diagonal and the size of the dot corresponds to the number of isolates obtained from patients of a certain age

born children of Mexican parents, presumed to have contracted TB by eating cheese made in Mexico from unpasteurized cow's milk [89]. The spoligotype pattern specific for the largest shared type in the outbreak, described by the 264073777777600 octal code, was found in almost half of the *M. bovis* strains isolated from the patients in our database.

Concomitant analysis of Fig. 5.3 and Fig. 5.4 allows making some interesting observations. For example, family 34 contains two elderly US-born patients of almost the same age. They are in fact infected with strains bearing an identical spoligotype (data not shown), which makes it highly probable that these two cases are involved in the same chain of recent transmission. Other explanations are possible, but these belong to the realm of epidemiologists. The X3 family included a group of mainly Caribbean immigrants aged over 40 years that have been in the US for at least 20 years. In this family, 12 persons shared the same spoligotyping pattern, which might be indicative of a recent outbreak among these long-time US residents. Alternatively, they all could have acquired the infection abroad and carried it for a long time before developing the active disease. The LAM4 family, comprising largely US-born patients, obviously manifests the spread of TB within NYC. The persons in this family should be an easier target of TB control practices; it has been demonstrated that US-born persons are more amenable to TB control measures, whereas dealing with infection among foreign-born individuals, especially when they carry latent infection, requires different measures [45].

The age of non-US-born patients, especially those born in countries with high incidence rates of TB, was skewed by the age at which the patients immigrate to the US, and may not reflect a real trend in the dissemination of TB. Figure 5.5 depicts the age distribution in foreign-born TB patients. The median and average ages are 38 and 42, respectively. It is apparent that age does not vary from family to family as much as in US-born patients. Again, the *M. bovis*-BCG family comprises the youngest patients.

Different research groups consistently observed that among TB patients male patients significantly prevailed over females [80, 108, 136]. Salihu et al. (2001) showed that males are at approximately twice the risk for the diseases than females

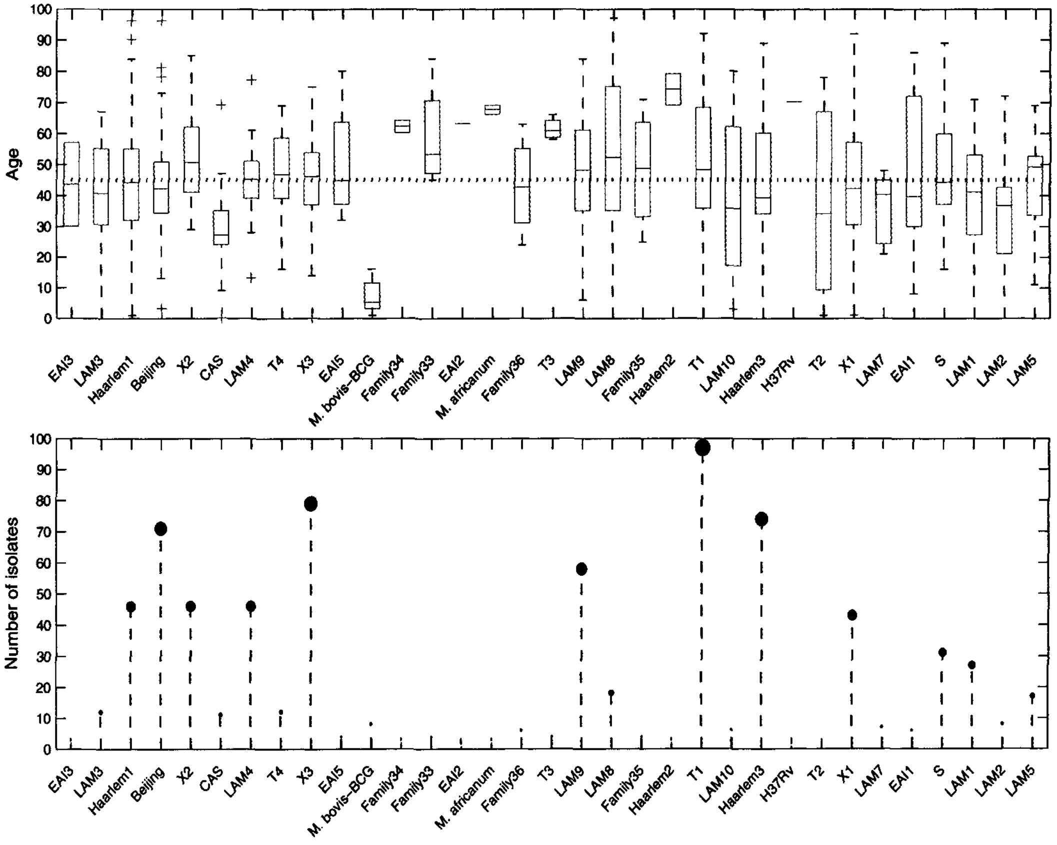


Figure 5.4: Top: Box plot of age at TB diagnosis of US-born patients by MTC strain families. The horizontal dotted line indicates median age. Bottom: Box plot displaying an approximate size of each of the families

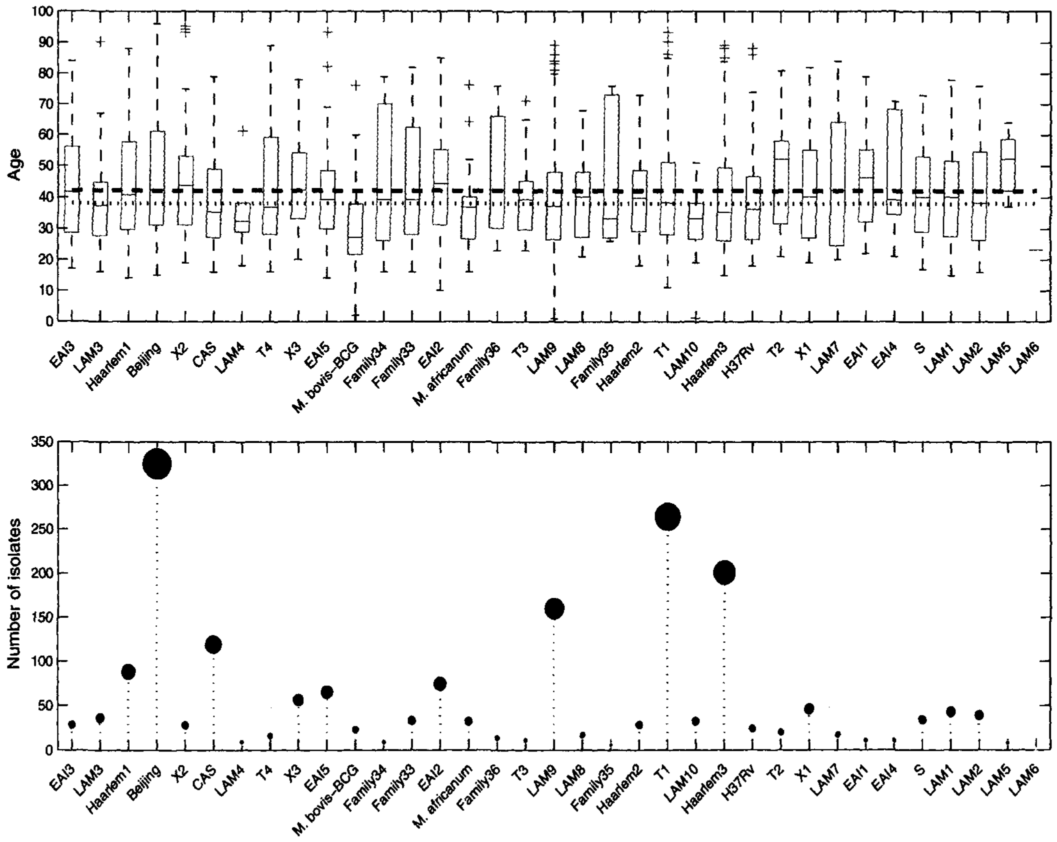


Figure 5.5: Top: Box plot of age at TB diagnosis of non-US-born TB patients by MTC strain families. The horizontal dotted and dashed lines indicate median and average age, respectively. Bottom: Box plot indicating an approximate size of each of the families

[108]. Overall, the NYC TB patient demographic data were consistent with this finding. Among the identified strain families, *M. bovis*-BCG is of interest in the context of patient gender distribution since it contained approximately equal number of persons of both genders. This reflected the fact that the majority of persons infected by strains from this family were young and was in accordance with an observation that from birth and up to the age of 24, there is no difference in TB risk by gender [108].

Family 36 consisted of equal numbers of males and females. In three EAI families, 1, 2 and 4, as well as in the LAM7 family, the number of females was higher. In the EAI1 family, the majority of the foreign-born patients (8/11) from several geographic regions have been in the US for less than 20 years (see Fig. 5.3), which indicates that the infection had been acquired abroad. The same applies to the EAI4 family, in which eight out of 11 strains were obtained from Vietnamese TB patients. One of these eight has been in the country for almost 22 years; the rest, for less than 16. A similar situation was observed in the EAI2 family, in which 79% of the isolates (59/75) came from the Philippines, a high TB burden country; 57 of the 59 have been living in the US for less that 21 years. The LAM7 family contained spoligotypes very different from the family's prototype, the main two of which had octal codes 776160000000071 and 776177400000171, defined as rare shared types (ST) 105 and 106, respectively, in SpolDB3. In this family, 13 out of 17 foreign-born persons have been in the US for less than 20 years. The foreign-born TB patients in the described families most probably acquired their infections outside of the US; therefore, the gender distribution within these groups of patients reflected immigration statistics. In 2000, a total of 42,197 males and 46,699 females immigrated to NYC [103].

We have closely examined the trends in the MTC data on patients from four countries with high incidence rates of TB: China, India, Ecuador, and Mexico.

The immigrants from these countries contribute most to the NYC's TB morbidity. Therefore, these groups are of particular interest to epidemiologists at whose request we designed several interactive methods for studying closely the TB dynamics within the countries with a high prevalence of TB (data not shown). Figure

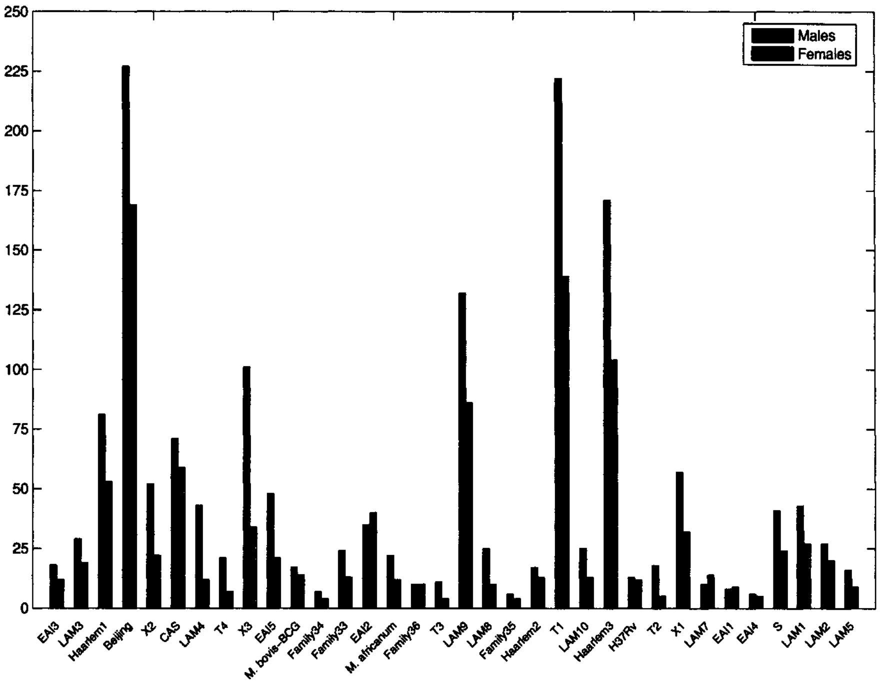


Figure 5.6: Gender distribution in TB patients from NYC database by MTC strain families

5.7 shows that the isolates from the four countries of interest occur in many of the spoligotyping families. Nevertheless, the majority of these isolates compose a few large families. Figure 5.8 demonstrates the distribution of the total number of isolates collected from patients born in each of the four countries.

The overwhelming majority of isolates from Chinese-born patients are allocated to three major families, Beijing, T, and Haarlem3; 15 isolates were identified as belonging to Haarleml family. All of these four families are prevalent worldwide, which has been demonstrated in Fig. 5.2. The ubiquitous Beijing type, which in 1995 was shown to characterize > 80% of isolates from China and to be a predominate strain in neighboring countries [150], has subsequently been associated with outbreaks or microepidemics worldwide [47]. Beijing/W strain is highly drug-resistant and caused large nosocomial outbreaks in NYC in the early 1990s [1]. The Haarlem type has recently been reported to have epidemic potential as well [81]. T, still poorly defined, is also a widespread family. It is noteworthy that many of the

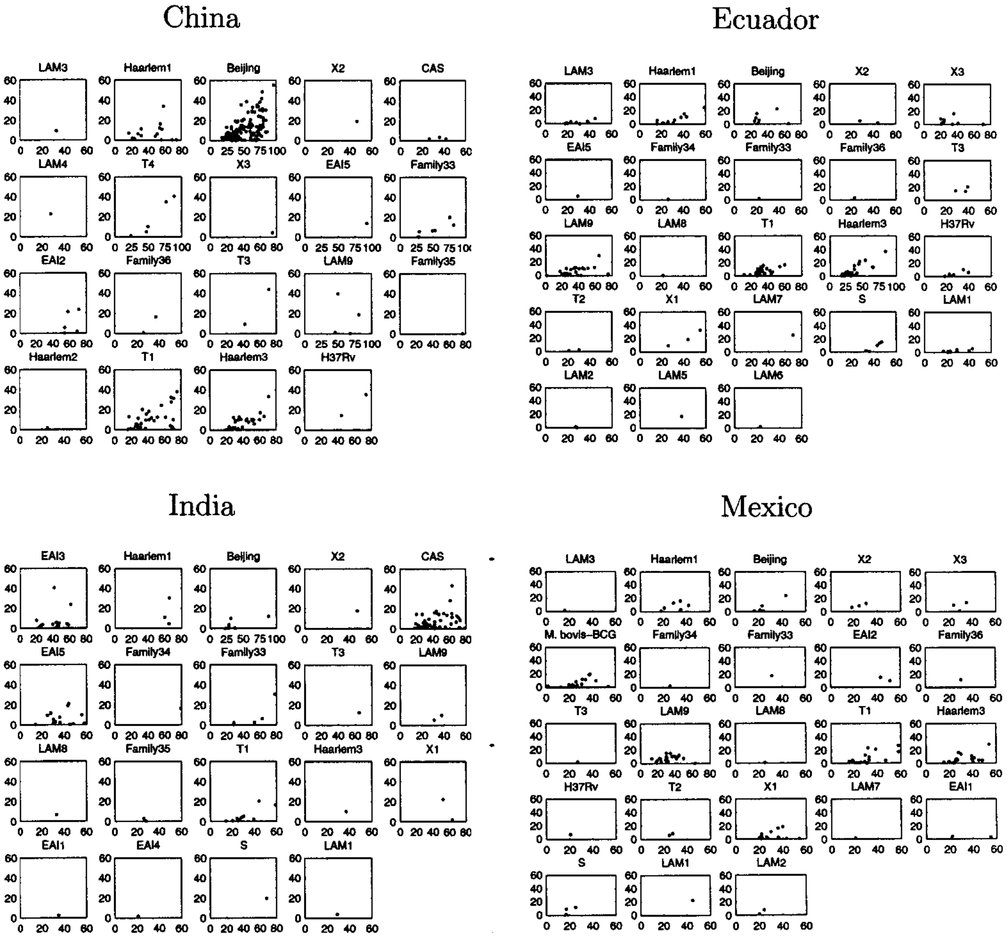


Figure 5.7: Time spent in the US by the onset of TB versus age of patients from the four countries with high incidence rates of TB, by the families identified using the SpolDB3-based model. Families are sorted by stability

Chinese-born patients infected with Beijing strains are aged over 40 and have been in the US for more than 20 years. This large family evidently includes TB cases resulting from both recently transmitted and reactivated latent infections. Figure 5.9 helps elucidate the duration of time spent in the US by the patients from the four countries of interest versus their age at the advent of TB. Interestingly, all except one patient in Haarlem3 had come to the US less than 20 years before the onset of TB, which indicated that their infections were imported.

The same four families containing isolates obtained from Chinese-born patients

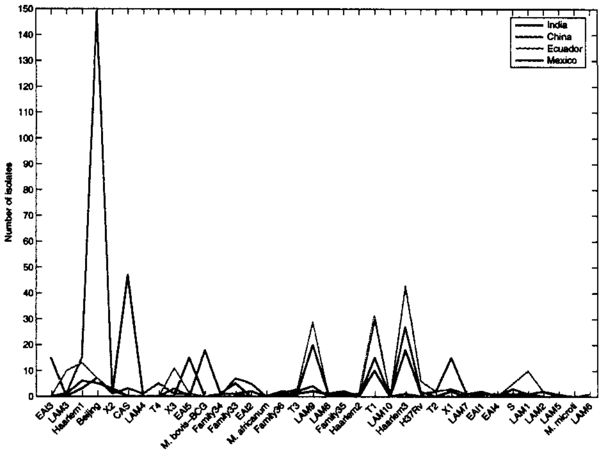


Figure 5.8: Total number of isolates obtained from patients born in India, China, Ecuador and Mexico, by the families identified using the SpolDB3-based model. Families are sorted by stability

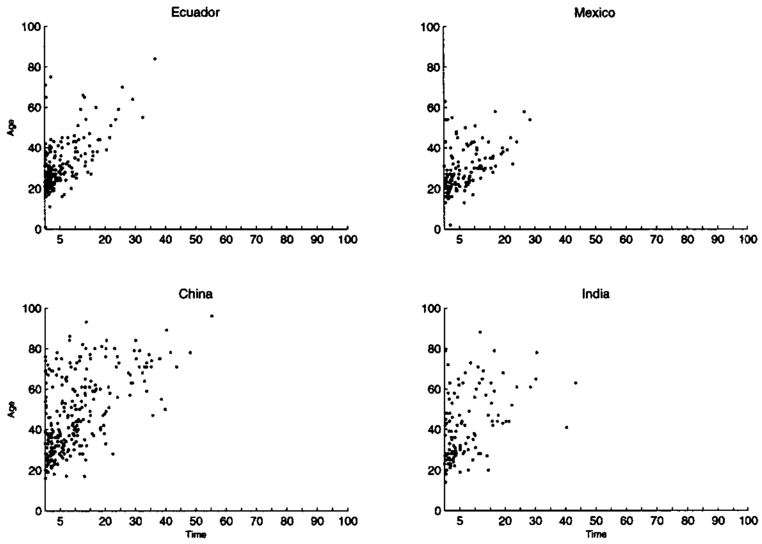


Figure 5.9: Distribution of age versus time in the US at the onset of TB in patients from the four countries with high incidence rates of TB

include the majority of the patients from Mexico. In 2005, 25% (1930/7656) of foreign-born cases in the US were reported in persons from Mexico [90]. Mexican-born patients from the NYC database are mostly aged between 20 and 40, with some outliers. The *M. bovis*-BCG, which was discussed above, requires special attention of TB controllers.

Isolates from patients born in Ecuador, a country with a very poor TB control [56], are found mostly in T, Haarleml and Haarlem3. Additionally, they total 30 in the LAM9 family, which can be considered a superfamily since its prototype is a parent of the prototypes for the LAM 1-8 families. Noticeably, Ecuadorian-born patients are, on average, relatively young, being aged below 40, and have not been in the US for a long time.

Isolates from patients originated in India, which accounts for 30% of TB cases worldwide, are prevailingly found in the CAS, EAI3, and EAI5 families. CAS has been shown to be a predominant genotype family in India [117]. In general, India is characterized by a TB dynamics different from that of other high burden countries.

The analysis of the countries with a high prevalence of TB reveals that most of the associated TB infections in NYC are caused by the isolates from a few ubiquitous families. We had at our disposal important but limited information on the foreign-born patients residing in NYC. When an isolate is classified into a prevalent worldwide genotype family, this fact alone is usually insufficient to provide clues into its origin to TB controllers. Additional genetic marker analysis and/or traditional epidemiological methods should be employed in this case. We can also conclude that in NYC preventive measures should be targeted toward recent immigrants.

In summary, our results demonstrated that the majority of TB cases reported in 2001-2004 in NYC occurred among foreign-born persons. This correlated with a prior observation that in NYC TB affects mostly the non-US-born population [136]. Our study corroborated a previous observation that US-born patients are more likely to belong to a cluster (in the epidemiological sense) than foreign-born persons [114, 116]; on average, non-US-born groups contained more shared types than the US-born patients within the same family. In addition, families CAS, EAI2, EAI5, and *M. africanum*, where foreign-born persons absolutely predominated, contained a significant number of unique isolates. Clustering is often assumed to indicate recent transmission of TB, while appearance of unique genotyping patterns suggests reactivation of latent infection [10]. Many different and often hard to track factors contribute to TB occurrences [61]; therefore, we believe that it is crucial to be very careful about making these assumptions based solely on genotyping data. For example, recent transmission may be underestimated when young patients are studied, and overestimated in the case of older individuals [152]. We can recognize some interesting patterns in the analyzed families and make suggestions on the associated TB dynamics. Thus, while the majority of the families contain mostly foreign-born patients, in several of the identified families US-born persons largely predominate. Clustered isolates within these families are most probably indicative of recent transmission. Unusually high number of unique isolates in family EAI5, taken together with the observation that these isolates were obtained from relatively young patients who originated in several different continents, strongly suggests multiple cases of reactivation of latent infection. Family *M. bovis*-BCG encompassed isolates from anomalously young persons and was discovered to result from a recent outbreak confirmed using genotyping data and by identified epidemiological links [89]. Average lower age of non-US-born patients than that of US-born persons, higher variations in shared types (data not shown) and a larger number of unique isolates among foreign-born persons suggest that most of these TB cases are due to imported infection. This is consistent with the previous finding that imported infection, either active or latent, is the cause of most TB cases among foreign-born persons in the US [144, 157]. The possibility of acquiring the infection in the US is much higher for non-US-born persons who have been in the country for over 20 years [116]. However, immigrants from countries with high rate of TB incidence can develop active disease even after having lived in the US for over 20 years [157].

We can conclude that patient data give an indispensable perspective on the spoligotyping families. Our results demonstrate the benefits of identifying the families as opposed to traditional epidemiological approach of searching for clusters of identical genotypes. We can detect unusual patterns of the patient data within the families; this would have been missed if we were to investigate our database as a whole. Our results allow us to reveal unsuspected trends in the infection spread and suggest possible scenarios of TB dissemination, thus directing efforts of TB control practices.

CHAPTER 6
Alternative methods of analysis of spoligotyping and patient data

The mixture model approach proved to be well suited for clustering MTC strain data. We also demonstrated that examining patient data in the context of spoligotyping families could be beneficial for TB epidemiology purposes. Here, we were interested in further improvement of our algorithm and discovering other approaches to model spoligotyping and patient data. Adopting existent techniques requires their thorough analysis. Modifications to known methods and new directions can be proposed based on the results of our experiments. In this chapter, we talk only about clustering as a statistical analysis technique of discovering natural groupings in the data, as opposed to identifying clusters of identical genotypes, routinely performed by epidemiologists.

We explored four alternative methods for analysis of the MTC strain data. First, we exploited principal component analysis, an essential multivariate data visualization technique, in an attempt to discover a suitable means for compact representation of our data and assessing clustering results. Second, we constructed joint mixture models for clustering combined spoligotyping and patient data. Third, we ascertained the possibility of using Bayesian networks for the mixture modeling. Lastly, a co-association matrix was created as a way of merging results of multiple clusterings, or a clustering ensemble. Visualization of the co-association matrix was shown to be a helpful tool in the exploratory analysis of the clustering results.

Validation of the results of our analyses presents a rather challenging problem, because the information available to date on the global spoligotyping families is limited to their prototypes, represented by the actual shared types [35]. Nevertheless, as more strains are being spoligotyped and new families are delineated within the global spoligotyping database, we become equipped with more tools to validate the results of the identification of MTC strain families. The fourth international spoligotyping database SpolDB4 has became publicly available recently [13].

SpolDB4 contains 62 prototypes for the MTC spoligotype families, thereby adding 26 new potentially phylogeographically-specific families. We discover that some of the families identified by our models correspond to the families that have been newly defined within the SpolDB4 database.

6.1 Principal component analysis

Principal component analysis (PCA) is a canonical statistical procedure widely used for dimensionality reduction of multivariate data. PCA transforms a number of correlated variables into a smaller number of uncorrelated variables called principal components. In essence, the method discovers the linear projections of the maximum variability in the data onto the lower dimensional subspace. The first principal component accounts for the greatest amount of the variation in the data; each succeeding component accounts for the next largest amount of variation and is independent of the preceding principal component. The maximum number of possible principal components is equal to the number of variables.

We have applied a logistic PCA to our spoligotyping data, which was developed especially for binary data [110, 140]. Logistic PCA is based on a multivariate generalization of the Bernoulli distribution. PCA computes a compact and optimal description of the data set; therefore, we were interested in examining the first few principal components constructed from the spoligotyping data.

Figure 6.1 shows the first three principal components selected among the original binary NYC spoligotyping data. The data points belonging to spoligotype families, or clusters, which PCA separates with varied success, are shown in different colors. All other clusters are superimposed on the axis origin and not shown. The families were identified using the SpolDB3-based model. The first three principal components capture the variation in our data associated with the major families that possess very distinct spoligotype signatures. The best separation was achieved for the Beijing, T1, Haarlem1 and CAS families. The isolates from the LAM families are plotted together since their binary prototypes are very similar. Besides, originally all of the LAM families were defined as one superfamily [113]. The logistic PCA does not account for the biological nature of spoligotypes; nevertheless, we

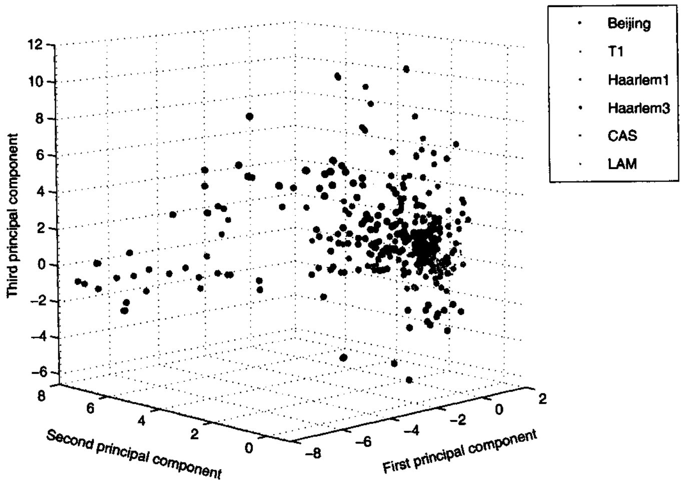
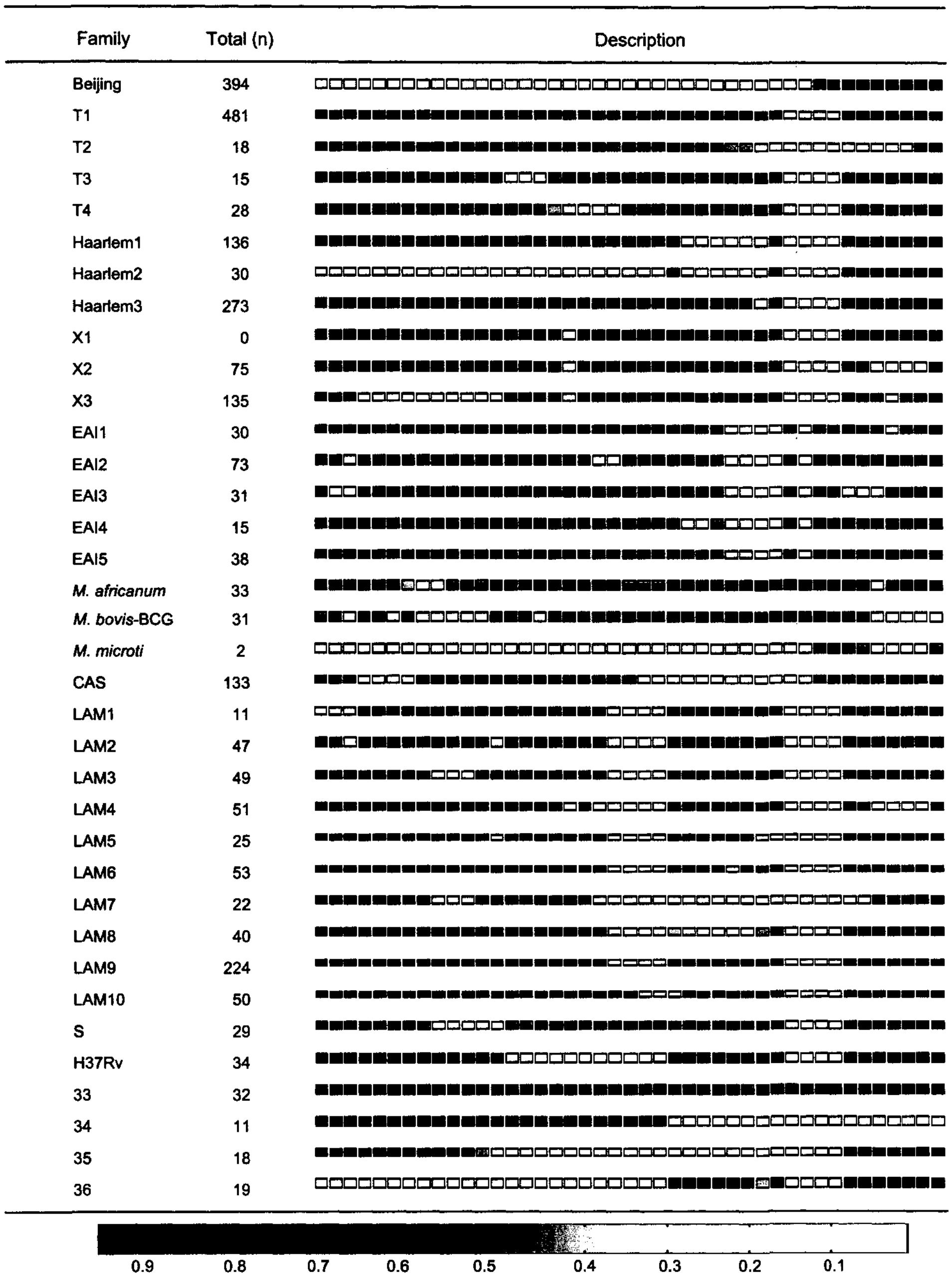


Figure 6.1: First three principal components found using the logistic PCA within the original spoligotype NYC data. Only the isolates belonging to the families that were separated are shown

can observe that it attains some success in extracting the directions of the highest variance within spoligotyping data and can be a useful tool for the visualization of the genotype families.

6.2 Joint mixture models

Analysis of the demographic data on patients infected with MTC isolates constituting the spoligotyping families proved to be valuable in discovering interesting TB trends. Ideally, the more information we have on TB cases, the more robust models we can potentially construct. However, this is only true if different types of data complement each other, together providing adequate information to discriminate well among MTC isolates. For example, the MIRU typing [132, 134] may not be able to provide sufficient discriminative power to resolve the genetic homogeneity of the Beijing family's spoligotyping patterns [65].



Family	Total (n)	Description
Beijing	394	
T1	481	
T2	18	
T3	15	
T4	28	
Haarlem1	136	
Haarlem2	30	
Haarlem3	273	
X1	0	
X2	75	
X3	135	
EAI1	30	
EAI2	73	
EAI3	31	
EAI4	15	
EAI5	38	
M. africanum	33	
M. bovis-BCG	31	
M. microti	2	
CAS	133	
LAM1	11	
LAM2	47	
LAM3	49	
LAM4	51	
LAM5	25	
LAM6	53	
LAM7	22	
LAM8	40	
LAM9	224	
LAM10	50	
S	29	
H37Rv	34	
33	32	
34	11	
35	18	
36	19	

Figure 6.2: Summary of the families identified within the NYC spoligotyping data using the SpolDB3-derived prototypes for model initialization. Probability of a spacer in Hidden Parent is represented by colored box; gradation of colors corresponds to probabilities of the spacer's presence: white indicates 0 and black indicates 1

Family	Total (n)	Description
Beijing	395	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■
T1	254	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
T2	30	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■▨□□□□■
T3	6	■■■■■■■■■■■■□□□□■■■■■■■■■■■■■■■■□□□□■■■■■■■
T4	29	■■■■■■■■■■■■■■■■■□□□□■■■■■■■■■■■□□□□■■■■■■■
Haarlem1	135	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■
Haarlem2	30	□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□□■□□□□■■■■■■■
Haarlem3	276	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■□□□□■■■■■■■
X1	232	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
X2	73	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■□□□□■
X3	135	■■■□□□□□□□□□□■■■■□■■■■■■■■■■■■■■□□□□■■■□■■■
EAI1	28	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■□■■■
EAI2	73	■■□■■■■■■■■■■■■■■■■□□■■■■■■■□□□□■□■■■■■■■■■
EAI3	32	■□□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■□□□■■■■
EAI4	12	■■■■■■■■■■■■■■■■■■■■■■■■■□□■□□□□■□■■■■■■■■■
EAI5	40	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■■■■■
M. africanum	33	■■■■■■▨□□■■■■■■■■■■■■■■■■■■■■■■■■■■■■□■■■■
M. bovis-BCG	31	■■□■■□■□□□□□■■■□■■■■■■■■■■■■■■■■■■■■■□□□□□
M. microti	1	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■□□□□□
CAS	133	■■■□□□□■■■■■■■■■■■■■■■■□□□□□□□□□□□□■■■■■■■■■
LAM1	3	□□□□■■■■■■■□□□□□■■■■□□□□■■■■□□□□□□□□□□□□□■■
LAM2	6	■■□■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■□□□□
LAM3	48	■■■■■■■■□□□■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■
LAM4	49	■■■■■■■■■■■■■■■■■□■□□□□□■■■■■■■■□□□□■■□□□□■
LAM5	26	■■■■■■■■■■■■□■■■■■■■□□□□■■■■■■□□□□□□■■■■■■■
LAM6	49	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■
LAM7	22	■■■■■■■■□□□■■■■■■■■□□□□□□□□□□□□□□□□□□□■■■■■
LAM8	39	■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□▨■□□□□■■■■■■■
LAM9	278	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■
LAM10	40	■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■■■□□□□■■■■■■■
S	29	■■■■■■■■□□□□□■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
H37Rv	34	■■■■■■■■■■■■■□□□□□□□□□□□■■■■■■■■□□□□■■■■■■■
33	37	■■■
34	11	■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□□
35	18	■■■■■■■■■■■▨□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■
36	19	□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■▨■□□□□■■■■■■■

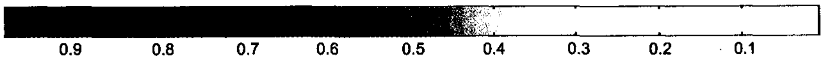


Figure 6.3: Summary of the families identified within the NYC spoligotyping and patient data using the SpolDB3-derived prototypes for model initialization. Probability of a spacer in Hidden Parent is represented by colored box; gradation of colors corresponds to probabilities of the spacer's presence: white indicates 0 and black indicates 1

In this section, we constructed joint mixture models integrating spoligotyping and patient data. The goal of the experiment was to assess whether the simultaneous modeling of spoligotyping and patient data helps in identifying the MTC strain families.

The NYC database contained information on the patient age, region of origin and time spent in the US at the advent of TB by the foreign-born patients. Here, we utilized these features in the context of a joint mixture model. The gender of each patient was also available; however, since this characteristic does not vary sufficiently in different families and in some families simply reflects the immigration statistics, we excluded it from the analysis. We did not have at our disposal a sufficient amount of any type of genotyping data other than spoligotyping.

We employed a joint mixture model that assumed that spoligotypes and the patient characteristics are independent of each other, conditioning on the family. Now, instead of finding the probability with which each spoligotype belongs to each of the families, we estimate this probability for each TB case associated with a particular patient. The probability of a case $\mathbf{y} = \{\mathbf{x}, \mathbf{a}, \mathbf{r}, \mathbf{t}\}$ being generated by a model C can be stated as follows:

$$P(\mathbf{y}) = \sum_{j=1}^{k} P(c_j) P(\mathbf{x}, \mathbf{a}, \mathbf{r}, \mathbf{t} | c_j, \theta_j), \tag{6.1}$$

where $\mathbf{x}$ is a spoligotyping pattern, $\mathbf{a}$ is the age, $\mathbf{r}$ is the region of patients' origin, and $\mathbf{t}$ is the time from the arrival in the US until the onset of TB for foreign-born patients. The time variable for US-born patients was equal to their age.

The age and time variables, being originally represented as continuous variables, were identically discretized into 7 following bins: $(0, 5]$, $(5, 20]$, $(20, 30]$, $(30, 40]$, $(40, 50]$, $(50, 60]$, and (> 60). For the time variable, the important points were 5 and 20 years spent in the US by the foreign-born patients. It is widely assumed by epidemiologists that if a foreign-born TB patient has spent less than 5 years in the US, he/she has most probably contracted the infection abroad. If this patient has been in the US for more than 20 years, the infection is considered to be acquired within the US. Persons born in countries with high incidence rates of TB may present an exception to this case: these TB patients were observed to have developed active TB from the acquired abroad latent infection even after having been in the US for more than 20 years. The geographical regions were divided into 8 bins ($m = 8$) as described in Chapter 5. Three isolates obtained from the Canadian-born patients were included in the group of isolates from the US-born patients.

The age and region variables were each equivalently modeled as a mixture of multinomial distributions wherein each mixture component corresponded to a family. Multinomial mixture models have been successfully used for text clustering [88], internet traffic clustering [63] and other problems. Multinomial distribution of a set of random variables $\{\mathbf{s_1}, \cdots, \mathbf{s_m}\}$ is a probability function

$$P(\mathbf{s_1} = s_1, \cdots, \mathbf{s_m} = s_m) = \frac{M!}{\prod_{h=1}^{m} s_h!} \prod_{h=1}^{m} \theta_h^{s_h}, \tag{6.2}$$

where s_h are nonnegative integers, $\sum_{h=1}^{m} x_h = M$, $\sum_{h=1}^{m} \theta_h = 1$, and $\theta_h > 0$. If $m = 2$, the multinomial distribution reduces to the binomial distribution. In the multinomial distributions for age and region, m =7 and 8, respectively. In our case $M = 1$, since each patient has a particular age and region of origin; therefore, the combinatorial term can be omitted. The probability that the model C has generated each of the age and region variables is defined as, for example, for the age:

$$P(\mathbf{a}) = \sum_{j=1}^{k} P(c_j) \prod_{h=1}^{m} pa_{jh}^{a_h}, \tag{6.3}$$

where pa_{jh} is the probability that the age of the patient falls into bin h and $a_h = 1$, since this is the number of different age variables in bin h.

We assume that the spoligotype, age, and region variables are mutually independent. The time variable, however, is dependent on both the age and region variables. Time is age-dependent because, if $\mathbf{a} = a$ and $\mathbf{t} = t$, $P(t > a) = 0$. The time that a foreign-born patient has spent in the US cannot be greater than his/her age. Time also depends on the region of origin because if a patient is born in the US, $P(t = a) = 1$. When the class is known, time depends only on age and region. The total log-likelihood function of the model parameters given the combined genotyping and patient data is defined as follows:

$$L(\Theta|X) = \sum_{i=1}^{n} log \sum_{j=1}^{k} P(c_j)P(\mathbf{x}_i|c_j)P(a_i|c_j)P(r_i|c_j)P(t|a_i, r_i), \tag{6.4}$$

where $P(\mathbf{x}_i|c_j)$ is defined in Eq. 4.4, and $P(a_i|c_j)$ and $P(r_i|c_j)$ are defined as in Eq. 6.3. When mixture component is known, the time variable depends only on the age and region variables. The $P(\mathbf{t} = t|\mathbf{a} = a, \mathbf{r} = r)$ was estimated from the NYC database. The spoligotyping data are modeled, as above, using the multivariate Bernoulli mixture model.

The EM algorithm was again used to estimate the optimizing parameters for now four types of data on each patient.

Algorithm 3: EM algorithm for joint mixture model

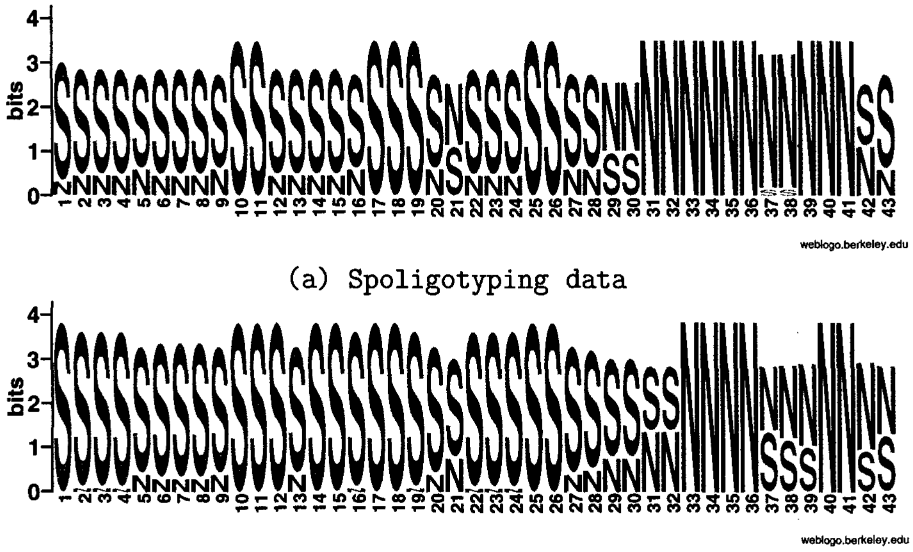
```
1. Choose initial parameter setting $\Theta' = \{P'(c_1), \cdots, P'(c_k), \theta'_1, \cdots, \theta'_k\}$.
   2. Repeat until convergence:
      E-step:
      $*E[z_{ij}] = \frac{P'(c_j)P(\mathbf{x}_i|c_j)P(a_i|c_j)P(r_i|c_j)P(t|a_i,r_i)}{\sum_{l=1}^{k} P'(c_l)P(\mathbf{x}_i|c_l)P(a_i|c_l)P(r_i|c_l)P(t_i|a_i,r_i)}$, $i = \{1, \cdots, n\}$, $j = \{1, \cdots, k\}$.

      M-step:
      $*P(c_j) = \frac{\sum_{i=1}^{n} E[z_{ij}]}{n}$, $j = \{1, \cdots, k\}$.
      $*p_{jd} = \frac{\sum_{i=1}^{n} E[z_{ij}]x_{id} - \sum_{i=1}^{n} E[z_{ij}]m_{10}}{\sum_{i=1}^{n} E[z_{ij}](m_{11} - m_{10})}$, $j = \{1, \cdots, k\}$, $d = \{1, \cdots, 43\}$.
      $*pa_{jh} = \frac{\sum_{i=1}^{n} E[z_{ij}]a_h}{n}$, $j = \{1, \cdots, k\}$, $h = \{1, \cdots, 7\}$.
      $*pr_{jh} = \frac{\sum_{i=1}^{n} E[z_{ij}]r_h}{n}$, $j = \{1, \cdots, k\}$, $h = \{1, \cdots, 8\}$.
      *Set $\Theta = \{P(c_1), \cdots, P(c_k), \theta_1, \cdots, \theta_k\}$.
```

The multivariate Bernoulli mixture model for the NYC spoligotyping data was initialized with the original SpolDB3-based prototypes. For the sake of consistency, we did not employ here the model trained on the New York State database, since these database did not contain patient information. It did include cumulative information on the countries of origin of patients associated with each shared type, which is available on our web site. This information however was not well suited for our method. Parameters for the multinomial probability distributions of the patient data variables, age and region, were initialized uniformly.

The results produced by the joint mixture model, which are shown in Fig. 6.3, were compared to the families identified by the model identically initialized with the SpolDB3-based prototypes and employing spoligotyping data only (Fig. 6.2). The comparison was performed by the visual inspection of the resulting families. The results showed that adding the patient data helped identify some of the families. We report only the results that we observed consistently, with different bin assignment and varied parameters m_{11} and m_{00}. These experiments are not shown because their validation belongs to future work. However, we are confident that the reported improvement in the quality of the three families, given our method and the NYC database, reflects the actual effect of adding demographic data to the model. Panel (b) of Fig. 6.4 shows that the patient data helped the identification of the T2 family. The spoligotypes with spacer 34 present were eliminated from the EAI4 family by incorporation of the patient data (Fig. 6.5). This is not surprising, since most of the isolates in this family have been obtained from Vietnam. In SpolDB4, the EAI4 was renamed to EAI4-VNM, where VNM stands for Vietnam [13]. Analogously, the LAM10 family, shown in Fig. 6.6, contains predominantly (34/40) African-born patients (Fig. 5.2), 28 out of which have been in the US for less than 5 years, and the other 6 - for less that 21.

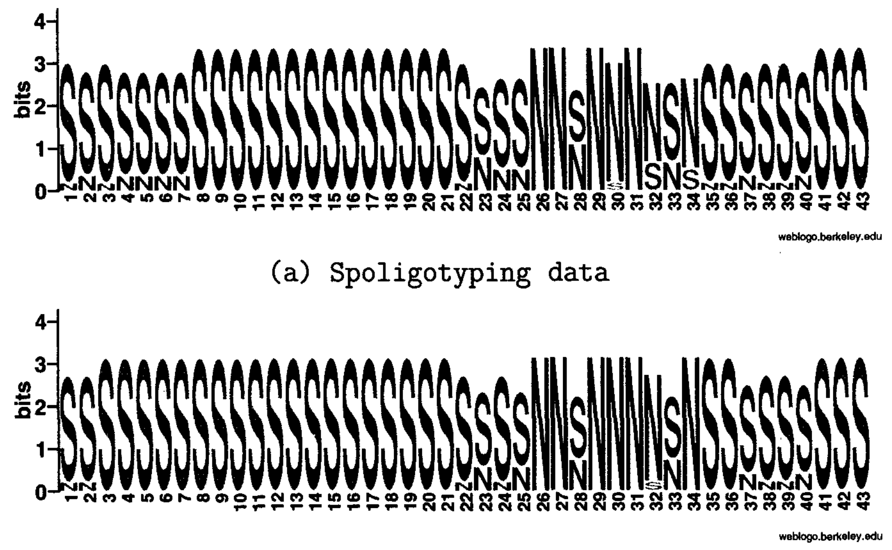
We also performed subclustering of the two large families identified by the 48-component RIM when applied to the New York State data (Fig. 4.11). According to the global databases SpolDB3 and 4, these two families, the children of the T1 and LAM9 prototypes, respectively, each contained isolates belonging to different genotyping families. We were interested to see whether the patient data could help identify smaller subfamilies within these diverse families. We carried out the subclustering using the NYC spoligotyping and patient MTC strain data. The 48-order RIM was first applied to NYC spoligotyping data, the two families, N24 and N19 were identified and then subclustered using: (1) spoligotyping data on these families, and (2) combined spoligotyping and patient data. Again, the Bernoulli mixture model was used for the experiment (1), and the mixture model of joint



(a) Spoligotyping data

(b) Spoligotyping and patient data

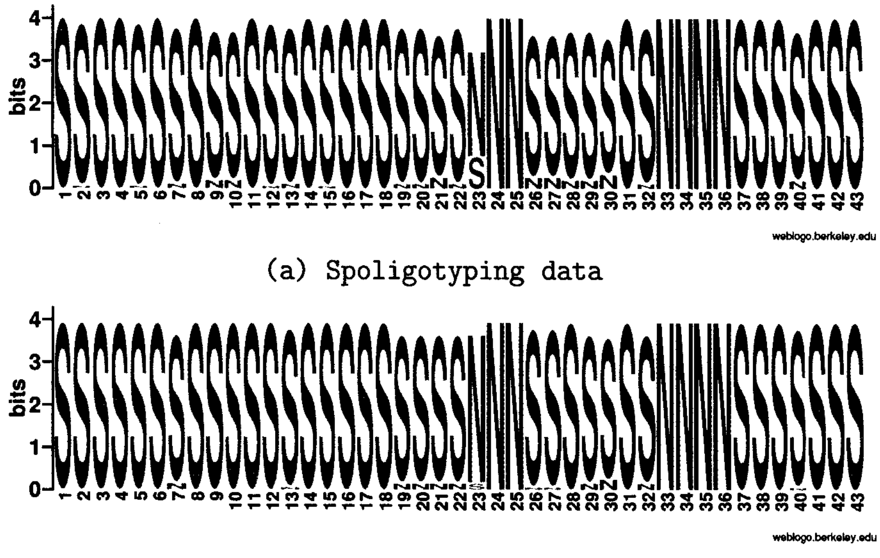
Figure 6.4: Sequence logo of the T2 family identified by the SpolDB3-based model within the NYC database



(a) Spoligotyping data

(b) Spoligotyping and patient data

Figure 6.5: Sequence logo of the EAI4 family identified by the SpolDB3-based model within the NYC database



(a) Spoligotyping data

(b) Spoligotyping and patient data

Figure 6.6: Sequence logo of the LAM10 family identified by the SpolDB3-based model within the NYC database

multivariate Bernoulli and multinomial distributions for the experiment (2).

The largest family N24 containing 700 isolates was subclustered using a 6-component model containing the SpolDB3-based prototypes for families T1, T2, T3, T4, Haarlem3 (H3), and X1. These six prototypes were chosen based on the visual inspection of the 700 isolates. Figure 6.7 shows the Hidden Parents for the six subfamilies identified within the superfamily N24. When the combined data were used as compared to spoligotyping data only, family T2 contained more isolates matching the prototype for this family (34 and 15, respectively). The patient data helped identify the T3 family. The T3 family in both cases contained two undesignated shared types, one having octal code 777703777760771, denoted as ubiquitous rare in SpolDB3, and its child with octal code 777703737760771 that is not present in SpolDB3. Also, when the NYC spoligotyping data only were used, the X1 family merged with the T1 isolates. Identification of the Haarlem3 family was hampered by adding the patient data; Fig. 5.2 and Fig. 5.3 indicate that the ubiquitous families T1 and Haarlem3 both contain isolates obtained from patients born in all of the geographic regions and of all of the age groups. The T family presents the most

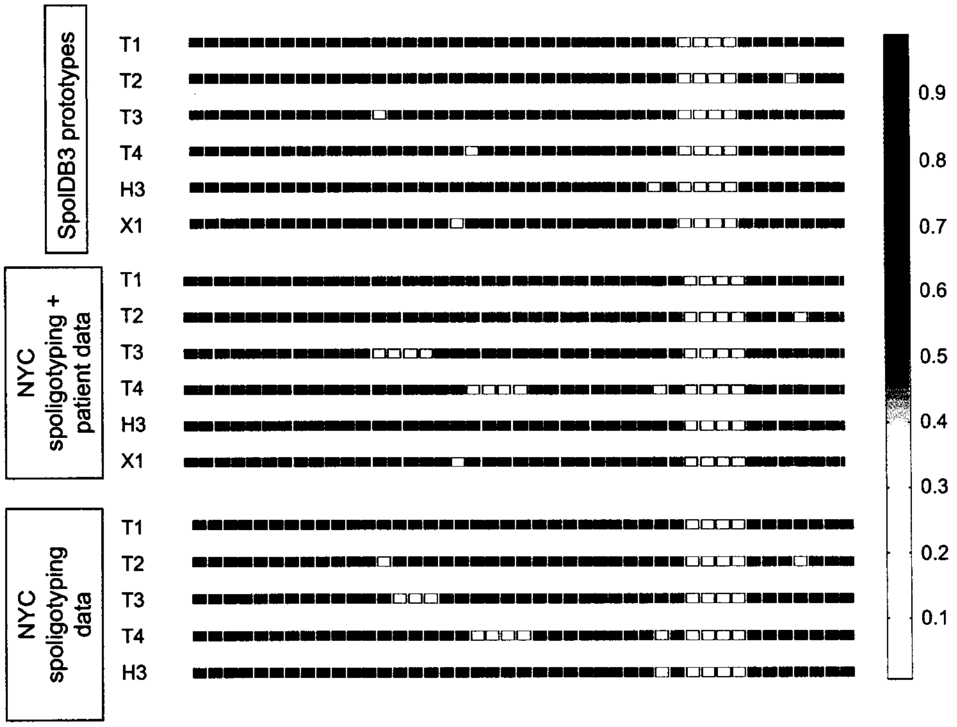


Figure 6.7: Results of subclustering on the "children of T1" superfamily (N24) identified within the NYC database using the 48-order RIM. The subclustering was carried out using the six SpolDB3-based prototypes

challenging case of identifying reliable families.

Subclustering of the N19 family, containing several subtypes of the LAM superfamily, was carried out using the SpolDB3-based prototypes for LAM1, LAM2, LAM5, LAM6, LAM8 and LAM9. These prototypes were chosen based on the visual analysis of the 336 spoligotypes in the N19 family identified within the NYC database. The Hidden Parents for the six LAM subtypes, trained on the two variants of the NYC data, along with the original SpolDB3-based prototypes, are shown in Fig. 6.8.

Integrating the spoligotyping and patient data on NYC TB patients facilitated the identification of the LAM2 and LAM5 families. The majority of the isolates in the LAM2 family were obtained from patient born in the Dominican Republic.

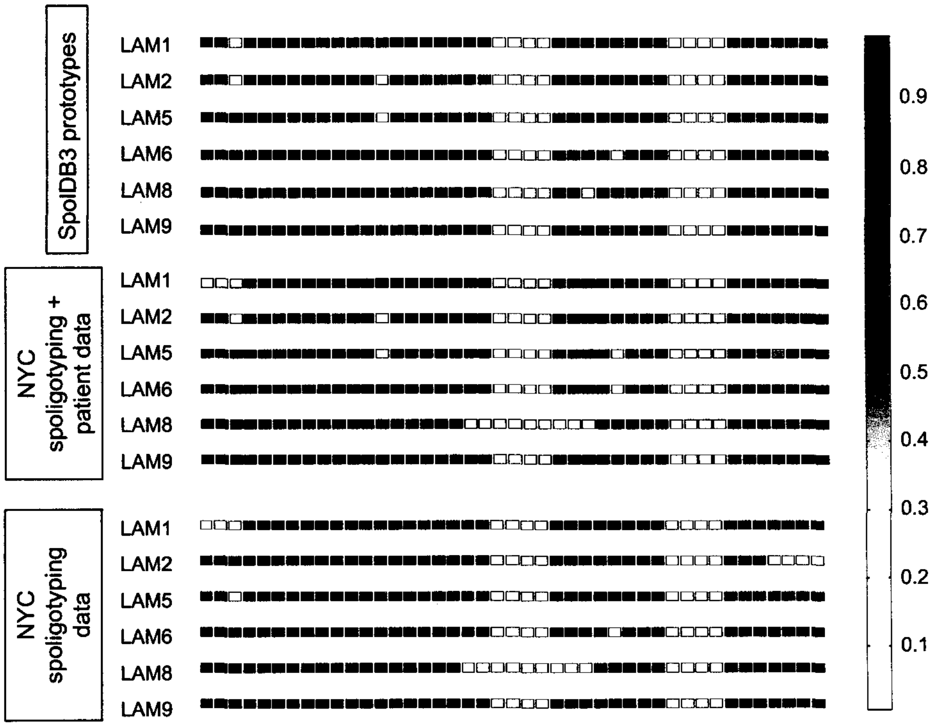


Figure 6.8: Results of subclustering on the "children of LAM9" family (N19) identified within the NYC database using the 48-order RIM. The subclustering was carried out using the six SpolDB3-based prototypes

Figure 6.8 demonstrates that the family identified within the combined data using the prototype for LAM5 presents an interesting case since its Hidden Parent looks like superimposed prototypes for LAM4, LAM5, and LAM6 (see Fig. 4.1). Four out of seven spoligotypes in this family have octal code 777737607560731. This shared type is documented in the National Quaternary Genotyping report (provided to us by NYC epidemiologists for internal use) and so far has been observed within the US in New York State only. Three of the patients infected with this shared type were born in Ecuador and one in Mali, the largest country in West Africa; the four patients are aged between 20 and 30 and have been in the US for less than 3 years. Of the other two persons in this family, one is from Mexico and one from Ecuador; both are also very recent immigrants. This information demonstrates that the strain is imported. When the NYC spoligotyping data only were used, the family initialized with the SpolDB3-based prototype for LAM5 resulted in a mixture of LAM1 and LAM2 isolates. In both of the experiments, the families identified using the LAM1 prototype had seven out of the total of 11 isolates bearing a spoligotype with octal code 077777607760771, which does not correspond exactly to any of the LAM families and is coined as rare in SpolDB3 [35]. The same applied to the LAM8 family that contained a shared type 777777000760771, a child of the prototype for this family.

It is obvious that the large superfamily LAM contains diverse assortment of spoligotypes which have not been yet divided into families. Patient data can help discriminate among these families. We summarize that subclustering can be employed as a simple and fast method to roughly elucidate the structure within complex families.

6.3 Co-association matrix as a method of combining multiple clusterings

In Chapter 4, we discussed that different families could be identified depending on the initial model parameters and the subsets of the data used. There obviously are many possible families that could be identified within our data. One can infer that if we are to analyze a global database having much higher spoligotyping pattern variability, the number of families that are potentially meaningful from computational and epidemiological points of view becomes very large.

Data clustering usually yields different results depending on the algorithm used, initial parameters, noise in the data and other factors. Clustering ensemble presents a way to exploit different partitions in the data. Various techniques can be used to generate different clusterings and the next challenging task could be to extract the consensus or the most optimal clustering partitions, which do not have to be the same.

A perturbation of the data set is often used to generate different clustering solutions and assess the stability of the clustering method. One way to perturb the data is to subsample the data points. While employing the MCCV approach [55], we subsampled the New York State database in order to estimate the optimal number of spoligotyping families. A theoretical basis for cluster validation by subsampling asserts that if the clusters capture real structure in the data than they are stable to minor perturbations. This approach is most appropriate when the data set possesses some redundancy. Another data perturbation approach is adding random noise to the data [7]. A spoligotype accompanied by the infected patient's data represents an actual TB case and cannot be "perturbed". Here we adopted an alternative approach of perturbation of the original SpolDB3-based prototypes by introducing randomness.

We investigated a powerful approach of combining multiple clustering partitions into a co-association, or co-occurrence, matrix that allows extracting different consensus partitions of the data [41, 57, 141]. This technique easily incorporates clusterings that differ in parameters, number of components, and that can be generated using different algorithms. The co-association matrix for clusterings of a data set of size n is a $n \times n$ matrix where each entry (i, j) contains the number of times the data points i and j occur in the same cluster, divided by the total number of clusterings. The cut-off value is used to join the data points into clusters. For example, if the cut-off value is 0.5, the data points that appear in the same cluster in 50% or more of the data partitions are placed in the same cluster in the consensus partition. Thus, different cut-off values allow us to extract different clusterings.

We analyzed the NYC database to exploit the applicability to our problem of the co-association matrix construction and processing. The SpolDB3-based prototypes were perturbed by introducing the 0.3% of randomness to each of the multivariate Bernoulli parameters for each of the model components. We generated 100 clusterings of the spoligotyping data. The same randomly perturbed prototypes were utilized to identify the families within the spoligotyping data merged with the patient data. The multinomial distributions for the age and region variables were initialized uniformly, without introducing additional randomness. Again, 100 clusterings were produced. The most stable clustering solution was recovered, based on the stability value calculated as described above. The two co-association matrices, one combining a clustering ensemble created using spoligotyping data only and another representing a clustering ensemble created using spoligotyping and patient data, were compared. The consensus clustering was also derived from the matrices.

It has been already discussed that as of today the pairwise distance between the spoligotyping patterns cannot be accurately devised, since the order and the number of DVRs that can be simultaneously lost or interrupted by insertion elements are unknown. The co-occurrence matrix can be used as a similarity matrix for the data points: the value of each matrix entry (i, j) indicates approximately how similar the two data points are. A convenient visualization toolkit, CLUSION, has been created to convert high-dimensional data into a perceptually more suitable format [130]. CLUSION reorders the data points according to a clustering solution λ so that the same cluster labels are contiguous, and then visualizes the resulting permuted similarity matrix. The original $n \times n$ similarity matrix $\mathbf{S}$ is permuted with an $n \times n$ permutation matrix $\mathbf{P}$, entries of which are defined as follows:

$$p_{ij} = \begin{cases} 1 & \text{if } j = \sum_{a=1}^{i} l_{a\lambda_i} + \sum_{l=1}^{\lambda_i - 1} n_l \\ 0 & \text{otherwise} \end{cases},$$

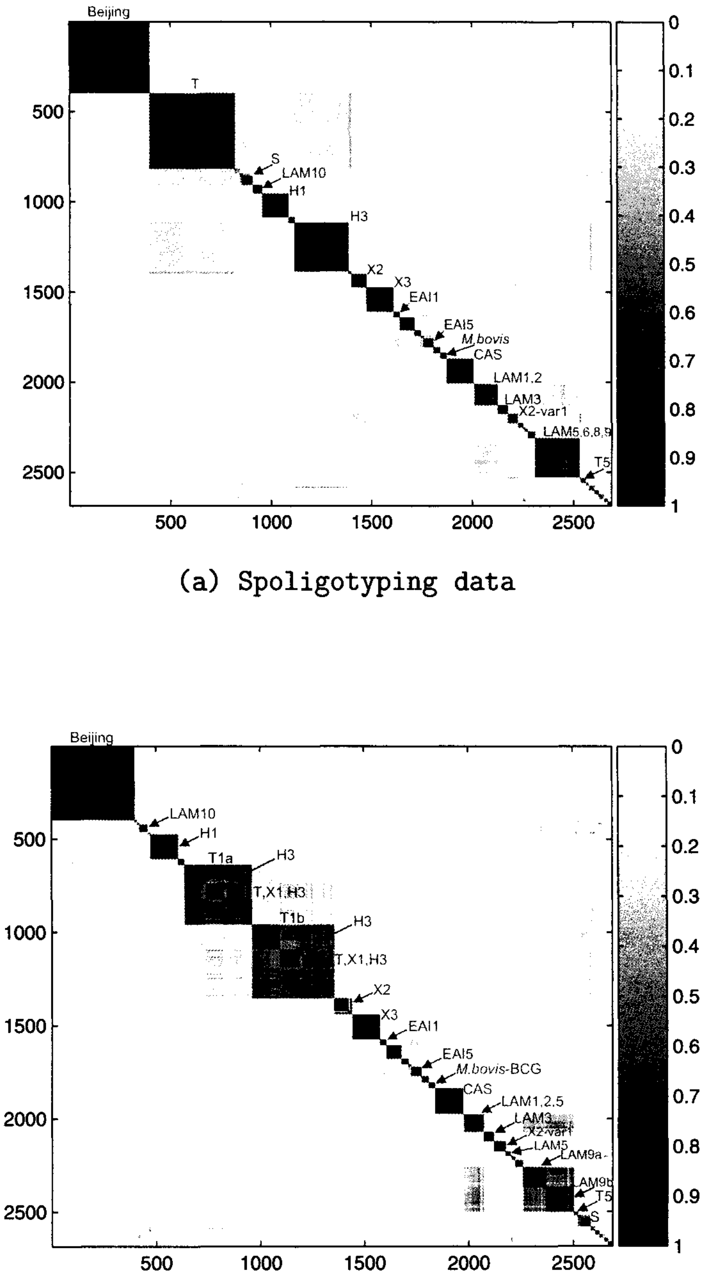
where l_{ij} is the entry of $\mathbf{L}$, a binary matrix representation of the cluster label vector λ. The indicator matrix $\mathbf{L}$ is defined by each entry l_{ij} as

$$l_{ij} = \begin{cases} 1 & \text{if } \lambda_i = j \\ 0 & \text{otherwise} \end{cases}.$$

The permuted similarity matrix $\mathbf{S}'$, the label vector λ' and data matrix $\mathbf{X}'$ are

$$\mathbf{S}' = \mathbf{PSP}', \quad \lambda' = \mathbf{P}\lambda, \quad \mathbf{X}' = \mathbf{PX}.$$

Seriation is defined as a reorganization of the dissimilarity values of a data matrix and has been used in various contexts, mainly in archaeology, psychology, and ecology (see, for example, [106]). It consists in bringing low dissimilarity (i.e. high similarity) values as close as possible to the main diagonal of the matrix. The seriated similarity matrix $\mathbf{S}'$ provides a relationship-centered view and is very useful



(a) Spoligotyping data

(b) Spoligotyping and patient data

Figure 6.9: Seriated similarity matrices. Cluster labels are produced by the most stable model among each of the 100 clusterings

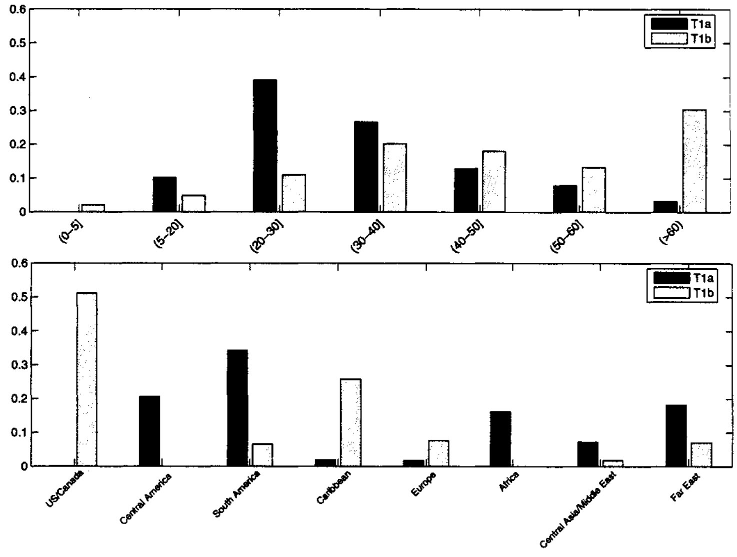


Figure 6.10: Difference in the parameters of the model components for the patient data in clusters T1a and T1b. Top: parameters for age. Bottom: parameters for regions

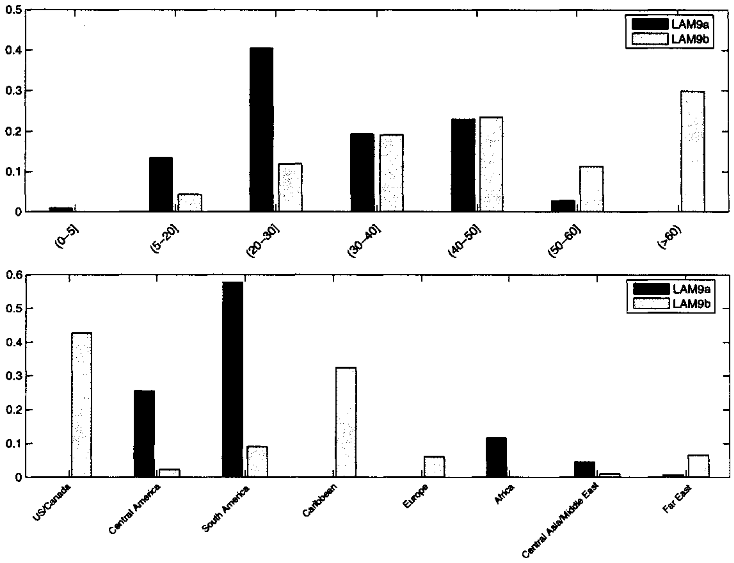


Figure 6.11: Difference in the parameters of the model components for the patient data in clusters LAM9a and LAM9b. Top: parameters for age. Bottom: parameters for regions

for visualization [130]. We adopted the underlying idea of the matrix seriation and reordered the data points within the co-association matrix according to the cluster labels produced by the most stable model. Figure 6.9 contains the two seriated matrices.

The gray-level images readily show the 36 identified families that initially had original order, as in SpolDB3 (see Fig. 4.1) [35]. The block-diagonal rectangular areas correspond to the clusters (families), and the intensity within each of the areas indicates the expected similarity within the cluster. In addition to examining the compact visual representation of the results of the two experiments (using the spoligotyping data only versus the spoligotyping data combined with the patient data), we manually explored each of the families in the two partitions. Since the models were initialized with the perturbed SpolDB3-based parameters, some of the identified families were different from those produced by the models initialized with the original SpolDB3 prototypes and described in Chapter 4.

The majority of the families previously suggested to be stable were resistant to the permutation of their prototypes: the EAI 2-4, *M. africanum*, CAS, LAM3, Haarlem1, 33 and 34 again proved to be well defined. *M. bovis*-BCG and Beijing were reproduced perfectly by both of the 100 clusterings. SpolDB4, wherein 62 lineage prototype patters were defined, provided new insights into the identified families [13]. Family 34 included type ST46 (ST stands for Shared Type) that was recognized as "undesignated" in SpolDB4. Similarly, family 36 contained ST4, which was termed LAM3/S in SpolDB4. Stable family LAM4 provides an example of how a new family found by our algorithm is defined within the global international database upon an increase in the size and diversity of the database. The Hidden Parent for this family was quite different from its initial prototype (Fig. 4.9); according to SpolDB4, this is the X2-variant1 family, or lineage. Analogously, the Hidden Parent for the stable T4 family, which was significantly different from the family' prototype, has been described as a new T5 family [13]. The prototype pattern for a newly defined in SpolDB4 T1-RUS2 family corresponds to the Hidden Parent for the family T3, identified in our data. Overall, despite these attempts to bring order into the T families, they are still not well defined [13].

The visualization of the seriated similarity matrices allows us to conveniently see which families were high-quality. The Beijing family, which comprises mainly one shared type ST1, is of very high quality and visualized as a dark homogeneous area at the top left corner of both matrices. The same applies to families Haarlem1 (H1), X2, X3, CAS, LAM3, EAI 1 and 5, and *M. bovis*-BCG, which are marked in Fig. 6.9. LAM10 is well defined using the randomized T4 prototype. Small families are not marked, but the identity of the dark block-diagonal areas corresponding to these families can be determining from Fig. 4.1. The families that are different from their SpolDB3 prototypes or receive new names upon the public appearance of SpolDB4, are marked. Stable families 33-36 are located in the lower right corner of the both panels of Fig. 6.9.

The seriated similarity matrices also illustrate which families presented a challenge for the algorithm. Dark off-diagonal regions suggest that the clusters in the corresponding rows and columns should be merged [130]. Panel (a) of Fig. 6.9 shows again that the T families are poorly defined: the block-diagonal area, adjacent to the Beijing family and marked as T, contains three subfamilies merged together. Family H3 overlaps with the big T cluster in some partitions. LAM families, except for LAM 3 and 10, constitute two clusters that are candidates for being merged into one cluster. Family S, resulted from the T3 prototype with added noise, overlaps with the T cluster. EAI5 isolates are occasionally interleaved with EAI1 isolates. A T5 family, newly defined in SpolDB4, was identified using the randomized LAM10 prototypes and is marked in Panel (b) of Fig. 6.9.

The clustering algorithm had difficulty in discriminating among H3 and T isolates when the spoligotyping data were modeled jointly with the patient data, which is demonstrated in Panel (b) of Fig. 6.9. This happens not only because the T and H3 families both have spoligotyping patterns that differ by the absence/presence of a very few spacers, H3 prototype being the child of T1 prototype, but also because the isolates from these families are found in patients of all age groups and born in many different countries. Nevertheless, the isolates characterized by the "children of T1" spoligotypes form two large separate clusters, marked as T1a and T1b in Panel (b) of Fig. 6.9. The H3 subcluster can be observed within each of these large clusters. The two T clusters are formed because the patient data associated with the clusters' isolates are different. The parameters for the clusters are shown in Fig. 6.10. For example, cluster denoted as T1a includes isolates only from foreign-born patients. Cluster T1b tends to include more isolates from older patients than cluster T1b.

One can observe that the clustering ensemble generated using the combined spoligotyping and patient data possesses more variability than the ensemble of clusterings of spoligotyping data. This is not surprising, because adding patient data only helps in the identification of several families. Panel (b) demonstrates variability in the identification of the T families, X1 overlaps with H3. On both figures, EAI5 spoligotypes are shown to occasionally be interleaved with EAI1.

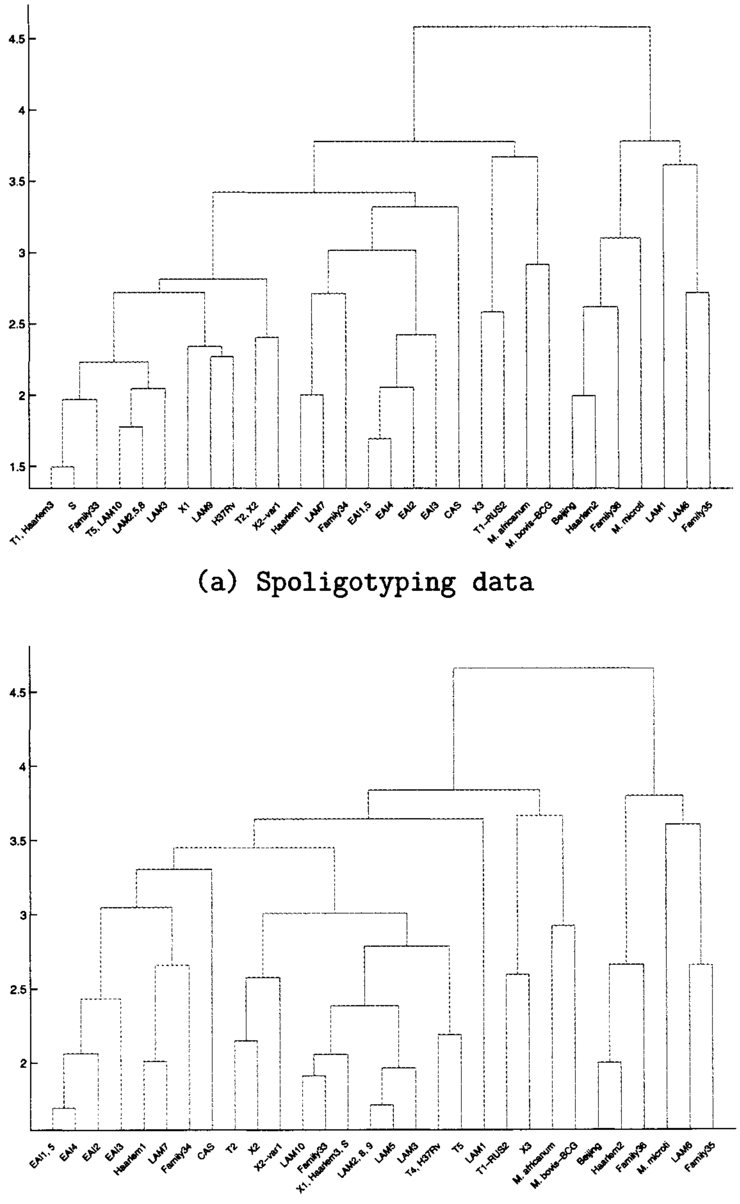
The examination of the areas corresponding to the families initialized with perturbed LAM prototypes gives interesting results. Panel (b) of Fig. 6.9 shows that the isolates bearing the "children of LAM9" spoligotypes make up two separate, but highly overlapping, families. Analogously to the case with the T families, we looked closely at the patient variables' parameters for these two LAM9 subclusters. Fig. 6.11 reveals that the LAM9a subcluster has a higher probability of including isolates from older patients than the LAM9a subcluster. LAM9a comprises isolates mostly from patients born in Central and South Americas, whereas LAM9a includes isolates obtained mainly from US- and Caribbean-born persons. The seriated matrix in Panel(b) of Fig. 6.9 suggests that the LAM9 and LAM 1, 2, and 5 clusters be merged. LAM5 forms a small separate family. Overall, combining patient and spoligotyping data seems to be useful in the identification of the LAM families. However, only LAM 3 and 10 can be consistently identified as distinct families when the SpolDB3 prototypes contain noise. These results once again signify that further efforts are required to develop robust algorithms for spoligotyping data clustering.

We also extracted the consensus partitions from the co-association matrices, using different cut-off values. The results are not presented here because the consensus clusterings do not provide any interesting results in addition to the analysis reported above. Essentially the same families are extracted at the cut-off value 0.5 as the families identified by the most stable model. When the cut-off values are higher, and therefore more clusters are extracted, the big families do not split into smaller ones, but instead more clusters containing one or very few isolates are formed. The PCA of the co-association matrices was performed. The results were very similar to those produced by the logistic PCA that was carried out using the original data; therefore, we do not provide here the PCA plots.

We performed hierarchical clustering of the Hidden Parents constituting the two most stable clusterings in an attempt to visualize the models from a perspective of the structure within the model. The euclidian distance and average linking were used, which approximated the unknown actual distances between spoligotyping patterns. Figure 6.12 depicts the dendrograms for the two models. We can observe that overall branch patterns are almost identical in both cases, which is not unpredicted, since we are looking at the genotyping data parameters only. The families with Hidden Parents having many spacers absent, such as Beijing, H2, family 35 and a few others, are separated from the rest of the families into the major left subtree. The major right subtree contains a separate branch for the *M. bovis*-BCG, *M. africanum*, X3, and T1-RUS2 (defined in SpolDB4) families. The clustering of the *M. bovis*-BCG and *M. africanum* types together have previously been observed [123]. The Hidden Parents for the EAI families form a separate subtree, with EAI 1 and 5 belonging to the same leaf node. In essence, we observe that the families whose Hidden Parents possess similar spoligotyping patterns, are naturally located close at the tree. This representation does not provide any insights into the evolutionary history of the MTC strain families; however, potentially it can be useful in the assessment of the overall relationships between existing spoligotyping families.

6.4 Bayesian networks

Our model with hidden variables that together are called Hidden Parent can be naturally represented as a directed acyclic graph (DAG). A DAG is a directed graph that contains no cycles. DAGs models represent conditional independencies among a set of random variables. In our case, an actual spoligotyping pattern is a child, or, to be more exact, a set of 43 children, and a Hidden Parent for the child's family is a parent. The parent directly influences the child by "producing" it. This



(a) Spoligotyping data

(b) Spoligotyping and patient data

Figure 6.12: Dendrogram of hierarchical clustering of the Hidden Parents comprising the two most stable models resulted from using the NYC MTC spoligotyping and merged spoligotyping and patient data, respectively

influence is represented by conditional probability. In the mixture model context, the child in independent of the mixture component if the parent is known.

In this section we describe our efforts in evaluating the potential of using DAGs in the form of Bayesian networks (BN) where each node corresponds to a variable. BNs allow us to model the causal relationships between the variables. They also readily adopt different structures. Moreover, BNs easily permit introduction of the interdependencies of the DVRs within the spoligotyping pattern.

6.4.1 Brief introduction to Bayesian networks

BNs are graphical models for representing the probabilistic relationships among a large number of variables and for doing probabilistic inference with those variables [97]. BNs were introduced by Kim and Pearl [68], Lauritzen and Speigelhalter [76], and others.

A BN for a set of variables $\mathbf{X} = \{X_1, \cdots, X_n\}$ is defined as a network structure S that encodes a set of conditional independencies among the variables in $\mathbf{X}$ and a set Θ of local probability distributions associated with each variable. The network structure is a DAG. Each node in the graph corresponds to a variable in $\mathbf{X}$. A directed arc from a variable X_j to another variable, X_i, indicates that X_j is a parent of X_i. The set of all the parents of the variable X_i is defined as $\mathbf{Pa}_i$. A conditional dependency links the child variable X_i to the set of parent variables $\mathbf{Pa}_i$ and is defined by the conditional distribution of the child variable given a configuration of the parent variables. In a BN, two children are independent of one another, given their parents, and are independent of all other nodes in the network given the nodes within their Markov blankets (parents, children, and parents of children). In other words, each node is conditionally independent of its non-descendants given its parents.

If we consider discrete variables only, then the joint probability distribution for X is given by:

$$P(\mathbf{x}) = \sum_{i=1}^{n} P(x_i|\mathbf{pa}_i), \tag{6.5}$$

where $x = \{x_1, \cdots, x_n\}$ is a value for $\mathbf{X}$, x_i is a value for X_i, and $\mathbf{pa}_i$ is a value for $\mathbf{Pa}_i$. If x_i has no parents, $P(x_i|\mathbf{pa}_i)$ stands for $P(x_i)$.

If we assume that the local probability distributions are defined by a finite set of parameters $\theta_s \in \Theta_s$, then the Eq. 6.5 can be represented as:

$$P(\mathbf{x}|\theta_s) = \sum_{i=1}^{n} P(x_i|\mathbf{pa}_i, \theta_s). \tag{6.6}$$

Let s^h denote the hypothesis that the "true" joint distribution for X can be represented by the DAG model s and has exactly the conditional independence assertions implied by s. We then obtain the following:

$$P(\mathbf{x}|\theta_s, s^h) = \sum_{i=1}^{n} P(x_i|\mathbf{pa}_i, \theta_s, s^h). \tag{6.7}$$

Many software packages were developed for learning both parameters and structure of BNs (for references, see [78]). Once a BN has been constructed from prior knowledge, data, or a combination of both, we can determine various probabilities of interest from the model. A BN for data X determines a joint probability distribution for X; therefore, in principle, we can use the BN to infer any probability of interest [53].

6.4.2 BNs for MTC strain data

6.4.2.1 Mixtures of DAGs

In essence, mixture models are one of the special cases of the general graphical model formalism [62]. The multivariate Bernoulli distribution can be readily represented by a DAG. Because we are dealing with a number of families, and, consequently, with a mixture model, we need to consider mixtures of DAG models (MDAG models) [138], where each DAG corresponds to a family. In this situation, the joint distribution of our data X is given by:

$$P(\mathbf{x}|\theta_s, s^h) = \sum_{c=1}^{|C|} P(c)P(\mathbf{x}|\theta_s, s_c^h). \tag{6.8}$$

This is a mixture of distributions determined by the component DAG models each having a mixture weight $P(c)$. The random variable C is hidden.

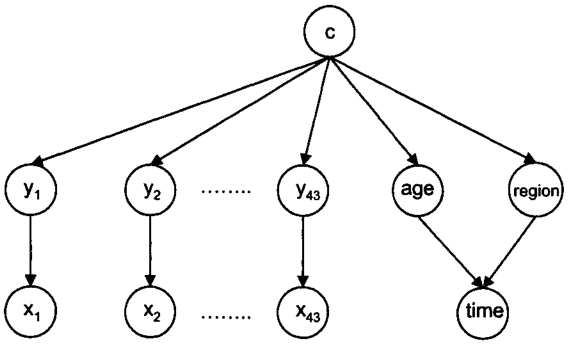


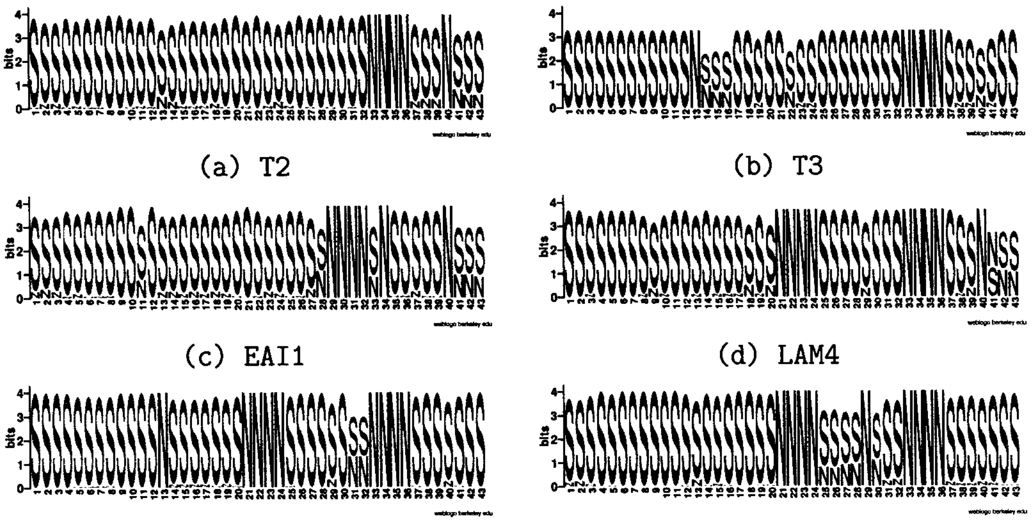
Figure 6.13: Schematic representation of the BN used to analyze MTC spoligotyping and patient data from NYC database. The top node is a class variable; the middle layer represents the Hidden Parent for the class $\{y_1, \cdots, y_{43}\}$ and the age and the region of origin of patients; and the bottom layer represents observed spoligotypes $\{x_1, \cdots, x_{43}\}$ and the time in the US spent by TB patients by the disease advent

We used the Matlab Bayes Net Toolbox (BNT) [94] to construct all of the BNs of interest, as well as to learn their parameters and infer required probabilities. Since BNs easily adopt different structures, at first, we constructed the model without Hidden Parents (data not shown). The only hidden variable each model component contained was the class variable. The spoligotype pattern was modeled as 43 observed variables. As expected, the absence of hidden variables severely deteriorated the performance of the model. Each of the identified families was a shared type that exactly matched the SpolDB3-based prototype used for the initialization of the corresponding component. Therefore, there was no variability within the families. One cluster contained an assortment of about 1500 spoligotypes that did not match exactly the original prototypes. These results confirm that hidden nodes are required to design robust models for clustering MTC genotype data.

Each DAG model included a hidden class variable and 43 hidden nodes each having a one-to-one correspondence to the Hidden Parent variables for the family.

Based on prior knowledge about the MTC strain data, we could construct a BN for both spoligotyping and patient data as shown in Fig. 6.13. One can immediately see that the graphical representation allows us to visualize the dependency of the time variable on the age and region variables. If the class is known, time depends only on age and region. These relationships were incorporated in the experiments with the joint mixture models described above. If we cluster spoligotyping data only, then the three nodes for the patient data are removed. The DAG in Fig. 6.13 shows a BN for each of the MTC strain families; the families have identical structures and differ only by the parameters. We therefore did not attempt to learn the structure of the DAGs that model different families.

We constructed the MDAG model using the BNT and inferred probabilities with which each of the data points belonged to each of the families. The junction tree engine was employed for the inference of all of the probabilities. We used the SpolDB3 prototypes to initialize the model's components; parameters for the class and patient variables were initialized uniformly. Each of the nodes was discrete and the patient were discretized as described above.



(a) T2

(b) T3

(c) EAI1

(d) LAM4

(e) LAM5

(f) LAM6

Figure 6.14: Logos of the families identified by the DAG mixture model schematically represented in Fig. 6.13 within the NYC spoligotyping and patient data

We have observed that the MDAG model allowed us to identify families that were much more consistent with their SpolDB3 prototypes as compared to the families distinguished by the Bernoulli mixture model, each conforming well to the hypothesis of losing DVRs in the course of the spoligotype's evolution. It was particularly remarkable in the case of the T families, all four of which matched their prototypes well, at the same time as possessed some reasonable variability of the patterns. Figure 6.14 shows the logos of the most prominent examples of the high quality families identified by the MDAG model. The families' logos can be compared to their respective prototypes shown in Fig. 4.1. Also, Panel (b) of Fig. 6.4 displays the logo representation of the T2 family identified by the joint mixture model. The comparison of this logo to the one depicted in Panel (a) of Fig. 6.14 clearly indicates that the BN is capable of identifying very high quality families. However, when the spoligotyping data alone were clustered by the MDAG model, the T families all merged into one.

Interestingly, when we tried to change the parameters m_{01} and m_{10}, which define the probabilities of losing and gaining a spacer from a parent, respectively, from the ones used by the EM algorithm described in Chapter 4, different results were observed. When $m_{01} = 10^{-2}$ and $m_{10} = 10^{-8}$, i.e. it is "harder" for a child spoligotype to either gain or lose a spacer, the families identified by the MDAG model within either the spoligotyping or the combined data were identical. The m_{01} and m_{10} parameters have been fixed throughout this study, but we have started to develop the models that would include an optimization step to learn these parameters for each family. Preliminary results of learning these parameters for each of the 43 positions of each of the model components were not satisfying.

We attempted to undertake the stability analysis of the MDAG models; however, the models were very sensitive to the perturbation of the initial parameters. One of the limitations of the BNs is that the results are highly dependent on the prior knowledge. The MDAG models initialized with the randomly perturbed SpolDB3 parameters did not produce nearly as high quality families as the models initialized with the best fitting the data, expert-defined parameters. The validation of the MTC data clustering using the MDAG models will be addressed in the future.

6.4.2.2 Mixtures of Hidden Markov Models

A Markov model [82] is a probabilistic process over a finite set, $\{x_1, \cdots, x_k\}$, usually called its states. The model describes at successive times the states of a system. At each of these times, the system may have changed from one state to another or stayed in the same state. The previous states are irrelevant for predicting the subsequent states, given knowledge of the current state. Each state transition generates a character from the alphabet of the process. The transition probabilities are defined as: $p_{ab} = P(x_t = b|x_{t-1} = a)$. Strictly speaking, this is a first-order discrete time Markov model where the probability at time t depends only on the states at time $\{t - 1\}$ [29]. In a Markov model, the state is directly visible to the observer; therefore, the state transition probabilities are the only parameters.

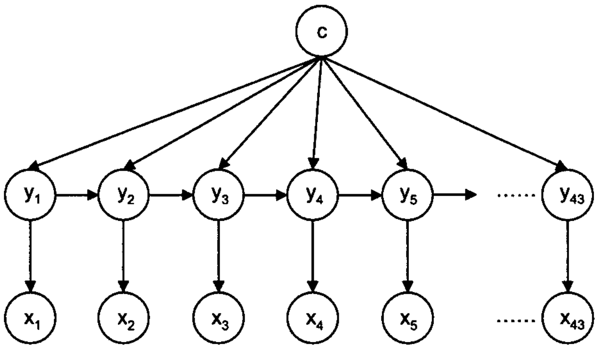


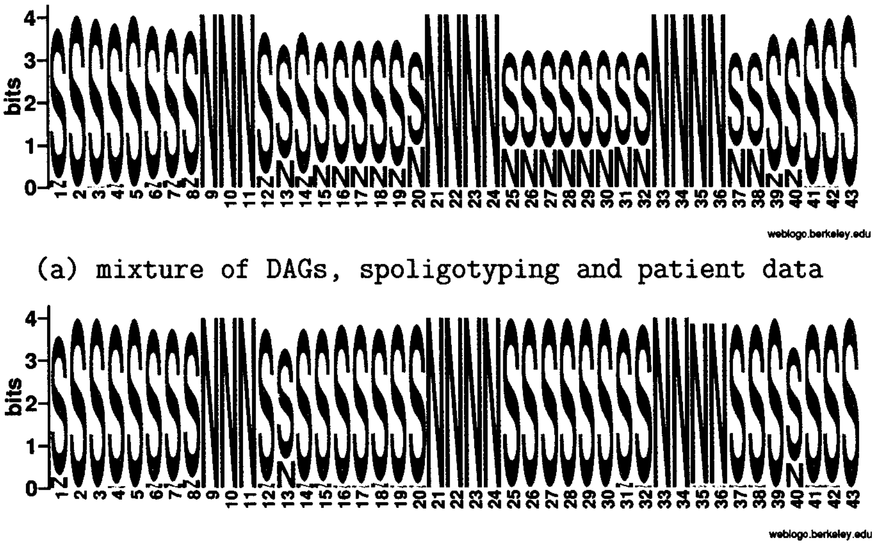
Figure 6.15: Schematic representation of the BN used to analyze MTC spoligotyping and patient data from NYC database. The top node is a class variable; the middle layer represents the Hidden Parent for the class $\{y_1, \cdots, y_{43}\}$, and the bottom layer represents observed spoligotypes $\{x_1, \cdots, x_{43}\}$

In most of the real-life systems, the states are not visible and the only measurable, or observed, variables are the emitted characters. A Markov model where the state sequence is hidden and each state has a probability distribution over the possible output characters, or tokens, is called a hidden Markov model (HMM) [102]. HMMs originally emerged in the domain of speech recognition and now are also widely used in biological sequence analysis and image recognition. The transition probabilities in HMMs are strictly causal, i.e. they depend only on previous states [29]. BNs are convenient tools for modeling these causal relationships.

We employed HMMs to determine whether introducing spacer interdependencies can help identify the MTC strain families. Although we have reasons to believe that DVRs within the DR region are dependent on each other, currently there is no direct evidence on which DVRs are interdependent. Nevertheless, it is reasonable to assume that adjacent DVRs interact, which is readily naturally modeled by HMMs where Hidden Parents are represented as hidden states. Figure 6.15 shows the BN that models spoligotype pattern as the first-order HMM. The initial parameters of the hidden states were the SpolDB3-derived prototypes modified to account for the possible values taken by the preceding state. We assumed that the transition probability matrix $P(x_t = a|x_{t-1} = b)$ was the same for all t, i.e. the probabilities are time-invariant. We conjectured that each succeeding state is more probable to emit a 0 if the preceding state had emitted a 0; thus, this allows us to integrate into the model the hypothesis asserting that multiple consecutive DVRs can be lost simultaneously .

Visual comparison of families identified by the MDAG and HMMs models using only the spoligotyping data permits to tentatively conclude that the HMMs are more suitable for clustering the MTC strain data. The families identified by the mixture of HMMs were almost identical to the families identified by the MDAG model among the spoligotyping and patient data. Some improvement in cluster quality was observed as well. Thus, the stable LAM3 family was more consistent with its prototype when the HMMs were utilized (Fig. 6.16). We intentionally delayed reporting the inconsistency in the content of the LAM3 family identified by the mixture of DAGs until this section.

The validation of the mixtures of HMMs was complicated for the same reason as the validation of the MDAG models, the high sensitivity of the algorithm to the permutation of the initial parameters. At this stage, we can suggest that the BNs should be used only when we are confident about the quality of the initial parameters. We can use alternative to the BNT software, or design our own computational



(a) mixture of DAGs, spoligotyping and patient data

(b) mixture of HMMs, spoligotyping data

Figure 6.16: Logo of the LAM3 family as identified by the MDAG and HMM mixture models in the NYC data

tool to cluster MTC data using BNs.

Conclusions and future work

The main contribution of this work consists in developing a new probabilistic modeling approach to the identification of MTC strain families based on spoligotyping data. The previous applications of data mining techniques to spoligotyping data did not take into account the biological nature of spoligotypes. We incorporated a widely assumed hypothesis on the spoligotypes's evolution into our model and showed that this allowed the identification of biologically meaningful families. TB epidemiologists, overwhelmed by the large volume and diversity of genotyping data, are sorely in need of efficient computational tools for processing the data. In collaboration with experts on TB molecular epidemiology, we initiated the project ultimately aimed at developing a robust computational system integrating heterogeneous data obtained by using different genetic markers. We concentrated on the analysis of spoligotypes because these data are fast and inexpensive to obtain, which is crucial for TB control programs, particularly in underdeveloped countries with high TB incidence rates.

Our results, based on the spoligotype analysis using mixture models, confirmed the reliability of the MTC strain families which had been previously defined empirically. Our results also identified certain new families of potential epidemiological value. We give examples of how the identified families can be used to examine trends in patient demographic data. Joint mixture modeling experiments further demonstrated the usefulness of supplementing genotype data with patient information. We also speculate that many epidemiologically meaningful spoligotyping families exist, which are yet to be recognized.

Several promising future directions are suggested. We explored our data in a graphical modeling framework. BNs were shown to identify the MTC strain families well, but were highly dependent on initial parameters. HMMs as an alternative method of modeling MTC strain data proved to be well suited for spoligotyping data. However, further work is required to develop validation techniques for MTC data clustering using graphical models.

Sparse knowledge of the evolution of spoligotypes is a major limiting factor in the designing of efficient models. Extensive simulation experiments are required. Also, we anticipate that the rapid advances in molecular techniques will provide us with biological information essential to the design of robust algorithms for global TB data analysis. In addition to improving models for spoligotyping and patient data, future work will concentrate on developing probabilistic models for merging other types of genotyping data with traditional epidemiological data. The first step in this direction will be to incorporate into our models the MTC strain data generated using typing methods such as mycobacterial interspersed repetitive units [84, 132, 133] and IS*6110*-based restriction fragment length polymorphism [146, 147].

Ultimately, our goal is to promote active and mutually beneficial collaboration aimed towards control of infectious diseases, using molecular methods of analysis, among TB controllers and biological and computational scientists. To this end, we have made it possible for users to submit their data to SPOTCLUST (http://www.rpi.edu/~bennek/EpiResearch). We hope that this will generate comments and suggestions from scientists with spoligotyping data of their own.

BIBLIOGRAPHY

[1] T.B. Agerton, S.E. Valway, R.J. Blinkhorn, K.L. Shilkret, R. Reves, and W.W. Schluter et al. Spread of strain W, a highly drug-resistant strain of *Mycobacterium tuberculosis*, across the United States. *Clin Infect Dis*, 29:85–92, 1999.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[3] A. Aranaz, B. Romero, N. Montero, J. Alvarez, J. Bezos, L. de Juan, A. Mateos, and L. Dominguez. Spoligotyping profile change caused by deletion of a direct variable repeat in a *Mycobacterium tuberculosis* isogenic laboratory strain. *J Clin Microbiol*, 42:5388–5391, 2004.

[4] L. Baker, T. Brown, M.C. Maiden, and F. Drobniewski. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis. Emerg Infect Dis*, 10:1568–1577, 2004.

[5] L.D. Baker and A.K. McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR-98)*, 1998.

[6] J.D. Banfield and A.E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[7] A. Ben-Hur and I. Guyon. Detecting stable clusters using principal component analysis. *Methods Mol Biol*, 224:159–182, 2003.

[8] P.J. Bifani, B. Mathema, N.E. Kurepina, and B.N. Kreiswirth. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. *Trends Microbiol*, 10:45–52, 2002.

[9] C.R. Braden, J. T. Crawford, and B.A. Schable. Assessment of *Mycobacterium tuberculosis* genotyping in a large laboratory network. *Emerg Infect Dis*, 8:1210–1215, 2002.

[10] C.R. Braden, G.L. Templeton, M.D. Cave, S. Valway, I.M. Onorato, K.G. Castro, D. Moers, Z. Yang, W.W. Stead, and J.H. Bates. Interpretation of restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates from a state with a large rural population. *J Infect Dis*, 175:1446–1452, 1997.

[11] P. Bradley, U. Fayyad, and C. Reina. Scaling EM (Expectation-Maximization) clustering to large databases. Technical report, 1997.

[12] R. Brosch, S.V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L.M. Parsons, A.S. Pym, S. Samper, D. van Soolingen, and S.T. Cole. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci U S A*, 99:3684–3689, 2002.

[13] K. Brudey, J.R. Driscoll, L. Rigouts, W.M. Prodinger, A. Gori, S.A. M. Al-Hajoj, C. Allix, L. Aristimuno, J. Arora, V. Baumanis, L. Binder, P. Cafrune, A. Cataldi, S. Cheong, R. Diel, C. Ellermeier, J.T. Evans, M. Fauville-Dufaux, S. Ferdinand, D. Garcia de Viedma, C. Garzelli, L. Gazzola, H.M. Gomes, M.C. Gutierrez, P.M. Hawkey, P.D. van Helden, G.V. Kadival, B.N. Kreiswirth, K. Kremer, and M. Kubin et al. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (spolDB4) for classification, population genetics and epidemiology. *BMC Microbiology*, 6, 2006.

[14] K. Brudey, M. Gordon, P. Mostrm, L. Svensson, B. Jonsson, C. Sola, M. Ridell, and N. Rastogi. Molecular epidemiology of *Mycobacterium tuberculosis* in Western Sweden. *J Clin Microbiol*, 42:30463051, 2004.

[15] M.V. Burgos, J.C. Mendez, and W. Ribon. Molecular epidemiology of tuberculosis: methodology and applications. *Biomedica*, 24(Supp):188–201, 2004.

[16] M.A. Carreira-Perpinan and S. Renals. Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, 12:141–152, 1999.

[17] F. Chaves, Z. Yang, H. el Hajj, M. Alonso, W.J. Burman, K.D. Eisenach, F. Dronda, J.H. Bates, and M.D.Cave. Usefulness of the secondary probe pTBN12 in DNA fingerprinting of *Mycobacterium tuberculosis*. *J Clin Microbiol*, 34:1118–1123, 1996.

[18] S.T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, and B.G. Barrell. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393:537–544, 1998.

[19] L.S. Cowan, L. Mosher, L. Diem, J.P. Massey, and J.T. Crawford. Variable-number tandem repeat typing of *Mycobacterium tuberculosis* isolates with low copy numbers of IS*6110* by using mycobacterial interspersed repetitive units. *J Clin Microbiol*, 40:1592–1602, 2002.

[20] G.E. Crooks, G. Hon, J.M. Chandonia, and S.E. Brenner. A sequence logo generator. *Genome Research*, 14:1188–1190, 2004.

[21] J.W. Dale, H. Al-Ghusein, S. Al-Hashmi, P. Butcher, A.L. Dickens, F. Drobniewski, K.J. Forbes, S.H. Gillespie, D. Lamprecht, T.D. McHugh, R. Pitman, N. Rastogi, A.T. Smith, C. Sola, and H. Yesilkaya. Evolutionary relationships among strains of *Mycobacterium tuberculosis* with few copies of IS*6110*. *J Bacteriol*, 185:2555–2562, 2003.

[22] J.W. Dale, D. Brittain, A.A. Cataldi, D. Cousins, J.T. Crawford, J. Driscoll, H. Heersma, T. Lillebaek, T. Quitugua, N. Rastogi, R.A. Skuce, C. Sola, D. van Soolingen, and V. Vincent. Spacer oligonucleotide typing of bacteria of the *Mycobacterium tuberculosis* complex: recommendations for standard nomenclature. *Int J Tuber Lung Dis*, 5:216–219, 2001.

[23] M.H. DeGroot and M.J. Schervish. *Probability and Statistics, 3rd Ed.* Addison-Wesley, Carnegie-Mellon University, 2002.

[24] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B(39):1–38, 1977.

[25] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[26] J.R. Driscoll, P.J. Bifani, B. Mathema, M.A. McGarry, G.M. Zickas, B.N. Kreiswirth, and H.W Taber. Spoligologos: a bioinformatic approach to displaying and analyzing *Mycobacterium tuberculosis* data. *Emerg Infect Dis*, 8:1306–1309, 2002.

[27] J.R. Driscoll, P. Lee, R. Jovell, Y. Hale, and M. Salfinger. How and why we fingerprint tuberculosis. *RT, The Journal for Respiratory Care Practitioners*, Feb/March 2001, 2001.

[28] V. Duchene, S. Ferdinand, I. Filliol, J.F. Guegan, N. Rastogi, and C. Sola. Phylogenetic reconstruction of *Mycobacterium tuberculosis* within four settings of the Caribbean region: tree comparative analyse and first appraisal on their phylogeography. *Infect Genet Evol*, 4:5–14, 2004.

[29] R. Duda, P. Hart, and D. Stork. *Pattern classification, 2nd Ed.* Wiley, New York, 2004.

[30] B. Efron. The jackknife, the bootstrap and other resampling plans. In *Society of Industrial and Applied Mathematics CBMS-NSF Monographs,38*, 1982.

[31] B.S. Everitt and D.J. Hand. *Finite mixture distributions. Monographs on statistics and applied probability.* Chapman and Hall, London, 1981.

[32] Z. Fang, N. Morrison, B. Watt, C. Doig, and K.J. Forbes. IS*6110* transposition and evolutionary scenario of the Direct Repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. *J Bacteriol*, 180:2102–2109, 1998.

[33] J.S. Farris. Phylogenetic analysis under Dollo's law. *Systematic Zoology*, 26:77–88, 1977.

[34] S. Ferdinand, G. Valetudie, C. Sola, and N. Rastogi. Data mining of *Mycobacterium tuberculosis* complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families. *Res Microbiol*, 155:647–654, 2004.

[35] I. Filliol, J.R. Driscoll, D. van Soolingen, B.N. Kreiswirth, K.Kremer, G. Valtudie, D.D. Anh, R. Barlow, D. Banerjee, P.J. Bifani, K. Brudey, A. Cataldi, R.C. Robert, C. Cooksey, D.V. Cousins, J.W. Dale, O.A. Dellagostin, F. Drobniewski, G. Engelmann, S. Ferdinand, D. Gascoyne-Binzi, M. Gordon, C. Gutierrez, W.H. Haas, H. Heersma, G. Kllenius, E. Kassa-Kelembho, T. Koivula, H.M. Ly, A. Makristathis, C. Mammina, G. Martin, P. Mostrm, I. Mokrousov, V. Narbonne, O. Narvskaya, A. Nastasi, S.N. Niobe-Eyangoh, J.W. Pape, V. Rasolofo-Razanamparany, M. Ridell, M.L. Rossetti, F. Stauffer, P.N. Suffys, H. Takiff, J. Texier-Maugein, V. Vincent, J.H. de Waard, C. Sola, and N.Rastogi. Global distribution of *Mycobacterium tuberculosis* spoligotypes. *Emerg Infect Dis*, 8:1347–1349, 2002.

[36] I. Filliol, J.R. Driscoll, D. van Soolingen, B.N. Kreiswirth, K.Kremer, G. Valtudie, D.A. Dang, R. Barlow, D. Banerjee, P.J. Bifani, K. Brudey, A. Cataldi abd R.C. Cooksey, D.V. Cousinsand, J.W. Dale, O.A. Dellagostin, F. Drobniewski, G. Engelmann, S. Ferdinand, D. Gascoyne-Binzi, M. Gordon, M.C. Gutierrez, W.H. Haas, H. Heersma, E. Kassa-Kelembho, M.L. Ho, A. Makristathis, C. Mammina, G. Martin, P. Mostrom, I. Mokrousov, V. Narbonne, O. Narvskaya, A. Nastasi, S.N. Niobe-Eyangoh, J.W. Pape, V. Rasolofo-Razanamparany, M. Ridell, M.L. Rossetti, F. Stauffer, P.N. Suffys, H. Takiff, J. Texier-Maugein, V. Vincent, J.H. de Waard, C. Sola, and N. Rastogi. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. *J Clin Microbiol*, 41:1963–1970, 2003.

[37] R.D. Fleischmann, D. Alland, J.A. Eisen, L. Carpenter, O. White, J. Peter-

son, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, E. Hickey, J.F. Kolonay, W.C. Nelson, L.A. Umayam, M. Ermolaeva, S.L. Salzberg, A. Delcher, T. Utterback, J. Weidman, H. Khouri, J. Gill, A. Mikula, W. Bishai, W.R. Jacobs, J.C. Venter, and C.M. Fraser. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol*, 184:5479–5490, 2002.

[38] N. Fomukong, M. Beggs, H. el Hajj, G. Templeton, K. Eisenach, and M.D. Cave. Differences in the prevalence of IS*6110* insertion sites in *Mycobacterium tuberculosis* strains: low and high copy number of IS*6110*. *Tuber Lung Dis*, 78:109–116, 1997.

[39] C. Fraley. Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20:270281, 1999.

[40] C. Fraley and A.E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 41:578–588, 1998.

[41] A. Fred. Finding consistent clusters in data partitions. In *Josef Kittler and Fabio Roli, editors, Multiple Classifier Systems, volume LNCS 2096*, page 309318, 2001.

[42] R. Frothingham and W.A. Meeker-O'Connell. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiology*, 144:1189–1196, 1998.

[43] S. Gagneux, K. Deriemer, T. Van, M. Kato-Maeda, B.C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M.C. Gutierrez, M. Hilty, P.C. Hopewell, and P.M. Small. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A*, 103:2869–2873, 2006.

[44] T. Garnier, K. Eiglmeier, J.C. Camus, N. Medina, H. Mansoor, M. Pryor, S. Duthoy, S. Grondin, C. Lacroix, C. Monsempe, S. Simon, B. Harris, R. Atkin, J. Doggett, R. Mayes, L. Keating, P.R. Wheeler, J. Parkhill, B.J.

Bart, G. Barrell, S.T. Cole, S.V. Gordon, and R.G. Hewinson. The complete genome sequence of *Mycobacterium bovis*. *Proc Natl Acad Sci U S A*, 100:78777882, 2003.

[45] E. Geng, B.N. Kreiswirth, C. Driver, J. Li, J. Burzynski, P. DellaLatta, A. LaPaz, and N.W. Schluger. Changes in the transmission of tuberculosis in New York City from 1990 to 1999. *N Engl J Med*, 346:1453–1458, 2002.

[46] Z. Ghahramani and M. Jordan. Supervised learning from incomplete data via an EM approach. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6.*

[47] J.R. Glynn, J. Whiteley, P.J. Bifani, K. Kremer, and D. van Soolingen. Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. *Emerg Infect Dis*, 8:843–849, 2002.

[48] P.M. Groenen, A.E. Bunschoten, D. van Soolingen, and J. D. van Embden. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol Microbiol*, 10:1057–1065, 1993.

[49] I. Gut. Automation in genotyping of single nucleotide polymorphisms. *Hum Mutat*, 17:475–492, 2001.

[50] M.M. Gutacker, J.C. Smoot, C.A. Migliaccio, S.M. Ricklefs, S. Hua, D.V. Cousins, E.A. Graviss, E. Shashkina, B.N. Kreiswirth, and J.M. Musser. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms. Resolution of genetic relationships among closely related microbial strains. *Genetics*, 162:1533–1543, 2002.

[51] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *J Intell Inform Syst*, 17:107145, 2001.

[52] J. He, M. Lan, C.L. Tan, S.Y. Sung, and H.B. Low. Initialization of clusters refinement algorithms: a review and comparative study. In *Proceedings of the*

International Joint Conference on Neural Networks (IJCNN 2004), Budapest, Hungary, 2004.

[53] D. Heckerman. A tutorial on learning with Bayesian Networks. In M. Jordan, editor, *Learning in Graphical Models.*

[54] A.E. Hirsh, A.G. Tsolaki, K. DeRiemer, M.W. Feldman, and P.M. Small. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *PNAS*, 101:4871–4876, 2004.

[55] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *Proc Natl Acad Sci U S A*, 101:5249–5253, 2004.

[56] S. Houston. Tuberculosis control in Ecuador: Unforeseen problems, unanticipated strengths. *Can Respir J*, 11, 2004.

[57] K. Inoue and K. Urahama. Fuzzy clustering based on cooccurrence matrix and its application to data retrieval. *Electronics and Communications in Japan*, 84:10–19, 2001.

[58] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bull Soc Vaud Sci Nat*, 44:223–270, 1908.

[59] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, NJ, 1988.

[60] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 3:264–323, 1999.

[61] R.M. Jasmer, J.A. Hahn, P.M. Small, C.L. Daley, M.A. Behr, A.R. Moss, J.M. Creasman, G.F. Schecter, E.A. Paz, and P.C. Hopewell. A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991-1997. *Ann Intern Med*, 130:971–978, 1999.

[62] M.I. Jordan. *Learning in Graphical Models.* MIT Press, 1998.

[63] M. Jorgensen. Using multinomial mixture models to cluster internet traffic. *Australian and New Zealand Journal of Statistics*, 46:205–218, 2004.

[64] A. Juan, J. Garca-Hernndez, and E. Vidal. EM initialization for Bernoulli mixture learning. *SSPR/SPR*, pages 635–643, 2004.

[65] K.M. Kam, C.W. Yip, L.W. Tse, K.L. Wong, T.K. Lam, K. Kremer, B.K.Y. Au, and D. van Soolingen. Utility of mycobacterial interspersed repetitive unit typing for differentiating multidrug-resistant *Mycobacterium tuberculosis* isolates of the Beijing family. *J Clin Microbiol*, 43:306–313, 2005.

[66] J. Kamerbeek, L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal, and J. van Embden. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol*, 35:5907–914, 1997.

[67] M. Kato-Maeda, P.J. Bifani, B.N. Kreiswirth, and P.M. Small. The nature and consequence of genetic variability within *Mycobacterium tuberculosis. J Clin Invest*, 107:533–537, 2001.

[68] J.H. Kim and J. Pearl. CONVICE; a conversational inference consolidation engine. *IEEE Trans. on Systems, Man and Cybernetics*, 17:120–132, 1987.

[69] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proceedings of the 14th International Conference on Machine Learning, Nashville, Tennessee; 1997 July 8-12*, 1997.

[70] K. Kremer, D. van Soolingen, R. Frothingham, W.H. Haas, P.W. Hermans, C. Martin, P. Palittapongarnpim, B.B. Plikaytis, L.W. Riley, M.A. Yakrus, J.M. Musser, and J.D. van Embden. Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. *J Clin Microbiol*, 37:2607–2618, 1999.

[71] S. Kullback. *Information theory and statistics.* John Wiley and Sons, New York, 1959.

[72] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:7986, 1951.

[73] N.E. Kurepina, S. Sreevatsan, B.B. Plikaytis, P.J. Bifani, N.D. Connell, R.J. Donnelly, D. van Sooligen, J.M. Musser, and B.N. Kreiswirth. Characterization of the phylogenetic distribution and chromosomal insertion sites of five IS*6110* elements in *Mycobacterium tuberculosis*: non-random integration in the dnaA-dnaN region. *Tubercle and Lung Disease*, 79:31–42, 1998.

[74] D.A. Langan, J.W. Modestino, and J. Zhang. Cluster validation for unsupervised stochastic model-based image segmentation. *ICIP-II*, 94:197–201, 1998.

[75] T. Lange, M.L. Braun, V. Roth, and J.M. Buhmann. Stability-based model selection. In *Advances in Neural Information Processing Systems (NIPS 2002)*, 2002.

[76] S. Lauritzen and D. Speigelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Royal statistical Society B*, 50:157–224, 1988.

[77] E. Legrand, I. Filliol, C. Sola, and N. Rastogi. Use of spoligotyping to study the evolution of the Direct Repeat locus by IS*6110* transposition in *Mycobacterium tuberculosis*. *J Clin Microbiol*, 39:1595–1599, 2001.

[78] P. Leray and O. Franois. Bnt structure learning package : Documentation and experiments. Technical Report FRE CNRS 2645, Laboratoire PSI, St-Etienne du Rouvray Cedex, France, 2004.

[79] J.H. Li, C.R. Driver, S.S. Munsiff, R. Yip, and P.I. Fujiwara. Differential decline in tuberculosis incidence among US- and non-US-born persons in New York City. *Int J Tuberc Lung Dis*, 7:451–457, 2003.

[80] Z. Liu, K.L. Shilkret, J. Tranotti, C.G. Freund, and L. Finelli. Distinct trends in tuberculosis morbidity among foreign-born and US-born persons in New Jersey, 1986 through 1995. *Am J Public Health*, 88:1064–1067, 1998.

[81] H. Mardassi, A. Namouchi, R. Haltiti, M. Zarrouk, B. Mhenni, and A. Karboul et al. Tuberculosis due to resistant Haarlem strain, Tunisia. *Emerg Infect Dis*, 11, 2005.

[82] A.A. Markov. Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya (in Russian)*, 15:135–156, 1906.

[83] B. Mathema, P.J. Bifani, J. Driscoll, L. Steinlein, N. Kurepina, S.L. Moghazeh, E. Shashkina, S.A. Marras , S. Campbell, B. Mangura, K. Shilkret, J.T. Crawford, R. Frothingham, and B.N. Kreiswirth. Use of spoligotyping to study the evolution of the Direct Repeat locus by IS*6110* transposition in *Mycobacterium tuberculosis*. *J Infect Dis*, 185:641–649, 2002.

[84] E. Mazars, S. Lesjean, A.L. Banuls, M. Gilbert, V. Vincent, B. Gicquel, M. Tibayrenc, C. Locht, and P. Supply. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proc Natl Acad Sci U S A*, 98:1901–1906, 2001.

[85] A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. In *Proceedings of the 15th National Conference on Artificial Intelligence*, 1998.

[86] A. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *Proceedings of the 15th International Conference on Machine Learning (ICML-98)*, pages 359–367. Morgan Kaufmann, 1998.

[87] G. McLachlan and K. Basford. *Mixture models*. Marcel Dekker, New York, NY, 1988.

[88] M. Meila and D. Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42:929, 2001.

[89] MMWR. Human tuberculosis caused by *Mycobacterium bovis* - New York City, 2001- 2004. June 24, 2005/54(24):605–608.

[90] MMWR. Trends in tuberculosis - United States, 2005. March 24, 2006/55(11):305–308.

[91] I. Mokrousov, O. Narvskaya, E. Limeschenko, T. Otten, and B. Vyshnevskiy. Novel IS*6110* insertion sites in the direct repeat locus of *Mycobacterium tuberculosis* clinical strains from the St. Petersburg area of Russia and evolutionary and epidemiological considerations. *J Clin Microbiol*, 40:1504–1507, 2002.

[92] P.J. Moreno, P.P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. Technical Report HPL-2004-4, Cambridge Research Laboratory, HP Laboratories, Cambridge, UK, 2004.

[93] P. Mostrom, M. Gordon, C. Sola, M. Ridell, and N. Rastogi. Methods used in the molecular epidemiology of tuberculosis. *Clin Microbiol Infect*, 8:694–704, 2002.

[94] Kevin Murphy. The BayesNet Toolbox for Matlab. In *Computing Science and Statistics: Proceedings of Interface*, volume 33, 2001.

[95] J.M. Musser. Molecular population genetic analysis of emerged bacterial pathogens: selected insights. *Emerg Infect Dis*, 2:1–17, 1996.

[96] J.M. Musser, A. Amin, and S. Ramaswamy. Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. *Genetics*, 155:7–16, 2000.

[97] R. Neapolitan. *Learning Bayesian Networks.* Prentice Hall, Upper Saddle River, NJ, 2004.

[98] C. Olivier, F. Jouzel, and A. El Matouat. Choice of the number of component clusters in mixture models by information criteria. In *Proceedings of the Vision Interface Conference, Trois-Rivieres, Canada*, pages 74–81, 1999.

[99] L.M. Parsons, R. Brosch, S.T. Cole, A. Somoskovi, A. Loder, G. Bretzel, D. van Soolingen, and Y.M. Hale amd M. Salfinger. Rapid and simple ap-

proach for identification of *Mycobacterium tuberculosis* complex isolates by PCR-based genomic deletion analysis. *J Clin Microbiol*, 40:2339–2345, 2002.

[100] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, Santa Mateo, CA, 1988.

[101] S. Poulet and S.T. Cole. Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis. Arch Microbiol*, 163:87–95, 1995.

[102] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

[103] RAND. A non-profit institution, url: http://ny.rand.org.

[104] I. Rish, J. Hellerstein, and T.S. Jayram. An analysis of data characteristics that affect naive Bayes performance. Technical report, 2001.

[105] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society*, B(49):223–239, 1987.

[106] W.S. Robinson. A method for chronologically ordering archaeological deposits. *American Antiquity*, 16:293–301, 1951.

[107] V. Roth, T. Lange, M. Braun, and J. Buhmann. A resampling approach to cluster validation. In *Proceedings of the 15th Symposium in Computational Statistics (COMPSTAT2002), Berlin , Germany*, pages 123–128, Heidelberg, 2002. Physica-Verlag.

[108] H.M. Salihu, E. Naik, W.F. O'Brien, G. Dagne, R. Ratard, and T. Mason. Tuberculosis in North Carolina: trends across two decades, 1980-1999. *Emerg Infect Dis*, 7:570–574, 2001.

[109] E. Savine, R.M. Warren, D. Gian, N. Beyers, P.D. van Helden, C. Locht, and P. Supply. Stability of variable-number tandem repeats of mycobacterial

interspersed repetitive units from 12 loci in serial isolates of *Mycobacterium tuberculosis*. *J Clin Microbiol*, 40:4561–4566, 2002.

[110] A.I. Schein, L.K. Saul, and L.H. Ungar. A generalized linear model for principal component analysis of binary data. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West, FL*, 2003.

[111] T.D. Schneider and R.M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res*, 18:6097–6100, 1990.

[112] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[113] M. Sebban, I. Mokrousov, N. Rastogi, and C. Sola. A data-mining approach to spacer oligonucleotide typing of *Mycobacterium tuberculosis*. *Bioinformatics*, 18:235–243, 2002.

[114] A. Seidler, A. Nienhaus, and R. Diel. The transmission of tuberculosis in the light of new molecular biological approaches. *Occupational and Environmental Medicine*, 61:96–102, 2004.

[115] J. Shao. Linear model selection by cross-validation. *J Am Stat Assoc*, 88:486–494, 1993.

[116] S. Sharnprapai, A.C. Miller, R. Suruki, E. Corkren, S. Etkind, J. Driscoll, M. McGarry, and E. Nardell. Genotyping analyses of tuberculosis cases in U.S.- and foreign-born Massachusetts residents. *Emerg Infect Dis*, 8:1239–1245, 2002.

[117] U.B. Singh, N. Suresh, N.V. Bhanu, J. Arora, H. Pant, S. Sinha, R.C. Aggarwal, S. Singh, J.N. Pande, C. Sola, N. Rastogi, and P. Seth. Predominant tuberculosis spoligotypes, Delhi, India. *Emerg Infect Dis*, 10:1138–1142, 2004.

[118] P. Smyth. Clustering using Monte Carlo cross-validation. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, OR*, pages 126–133. AAAI Press, 1996.

[119] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10:63–72, 2000.

[120] H. Soini, X. Pan, L. Teeter, J.M. Musser, and E.A. Graviss. Transmission dynamics and molecular characterization of *Mycobacterium tuberculosis* isolates with low copy numbers of IS*6110*. *J Clin Microbiol*, 39:217–221, 2001.

[121] C. Sola, A. Devallois, L. Horgen, J. Maisetti, I. Filliol, E. Legrand, and N. Rastogi. Tuberculosis in the Caribbean: using spacer oligonucleotide typing to understand strain origin and transmission. *Emerg Infect Dis*, 5:404–414, 1999.

[122] C. Sola, S. Ferdinand, L.A. Sechi, S. Zanetti, D. Martial, C. Mammina, A. Nastasi, G. Fadda, and N. Rastogi. *Mycobacterium tuberculosis* molecular evolution in western Mediterranean islands of Sicily and Sardinia. *Infect Genet Evol*, 5:145–156, 2005.

[123] C. Sola, I. Filliol, M.C. Gutierrez, I. Mokrousov, V. Vincent, and N. Rastogi. Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives. *Emerg Infect Dis*, 7:390–396, 2001.

[124] C. Sola, I. Filliol, E. Legrand, S. Lesjean, C. Locht, P. Supply, and N. Rastogi. Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. *Infect Genet Evol*, 3:125–133, 2003.

[125] C. Sola, I. Filliol, E. Legrand, I. Mokrousov, and N. Rastogi. *Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS1081, IS*6110*, VNTR and DR-based spoligotyping suggests the existence of two new phylogeographical clades. *J Mol Evol*, 53:680–689, 2001.

[126] R.S. Spurgiesz, T.N. Quitugua, K.L. Smith, J. Schupp, E.G. Palmer, R.A. Cox, and P. Keim. Molecular typing of *Mycobacterium tuberculosis* by using nine novel variable-number tandem repeats across the Beijing family and low-copy-number IS*6110* isolates. *J Clin Microbiol*, 41:4224 - 4230, 2003.

[127] C. Sreevatsan, X. Pan, K.E. Stockbauer, N.D. Connell, B.N. Kreiswirth, T.S. Whittam, and J.M. Musser. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci U S A*, 94:9869–9874, 1997.

[128] W.W. Stead, K.D. Eisenach, M.D. Cave, M.L. Beggs, G.L. Templeton, C.O. Thoen, and J.H. Bates. When did *Mycobacterium tuberculosis* infection first occur in the New World? An important question with public health implications. *Am J Respir Crit Care Med*, 151:1267–1268, 1995.

[129] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2002.

[130] A. Strehl and J. Ghosh. Relationship-based clustering and visualization for high-dimensional data mining. *INFORMS Journal on Computing*, 15:208–230, 2003.

[131] C.A. Sugar and G.M. James. Finding the number of clusters in a dataset: An information-theoretic approach. *J Am Stat Assoc*, 98:750–763, 2003.

[132] P. Supply, S. Lesjean, E. Savine, K. Kremer, D. van Soolingen, and C. Locht. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. *J Clin Microbiol*, 39:3563–3571, 2001.

[133] P. Supply, J. Magdalena, S. Himpens, and C. Locht. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol*, 26:991–1003, 1997.

[134] P. Supply, E. Mazars, S. Lesjean, V. Vincent, B. Gicquel, and C. Locht. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol Microbiol*, 36:762–771, 2000.

[135] P. Supply, R.M. Warren, A.L. Bauls, S. Lesjean, G.D. van Der Spuy, L.A. Lewis, M. Tibayrenc, P.D. van Helden, and C. Locht. Linkage disequilibrium

between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Mol Microbiol*, 47:529–538, 2003.

[136] E.A. Talbot, M. Moore, E. McCray, and N.J. Binkin. Tuberculosis among foreign-born persons in the US, 1993-1998. *JAMA*, 284:2894–2900, 2000.

[137] M.M. Tanaka, P.M. Small, H. Salamon, and M.W. Feldman. The dynamics of repeated elements: Applications to the epidemiology of tuberculosis. *Proc Natl Acad Sci U S A*, 97:3532–3537, 2000.

[138] B. Thiesson, C. Meek, D.M. Chickering, and D. Heckerman. Learning mixtures of DAG models. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98), July 24-26, 1998, University of Wisconsin Business School, Madison, Wisconsin, USA*, pages 504–513. Morgan Kaufmann, 1998.

[139] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistical Society*, B(63):411–423, 2001.

[140] M. Tipping. Probabilistic visualization of high-dimensional binary data. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 592–598. MIT Press, 1999.

[141] A. Topchy, A.K. Jain, and W. Punch. Combining multiple weak clusterings. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03), Melbourne, Florida*, pages 331–338.

[142] A. Topchy, A.K. Jain, and W. Punch. A mixture model for clustering ensembles. In M.W. Berry, U. Dayal, C. Kamath, and D. Skillicorn, editors, *Proceedings of the 4th SIAM International Conference on Data Mining; Lake Buena Vista, Florida*, 2004.

[143] A. Topchy, M.H.C. Law, A.K. Jain, and A.L. Fred. Analysis of consensus partition in cluster ensemble. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04)*, pages 225–232.

[144] N.G. Tornieporth, Y. Ptachewich, N. Poltoratskaia, B.S. Ravi, M. Katapadi, J.J. Berger, and M. Dahdouh et al. Tuberculosis among foreign-born persons in New York City, 1992-1994: implications for tuberculosis control. *Int J Tuberc Lung Dis*, 1:528–535, 1997.

[145] Y. Tsuruoka and J. Tsujii. Training a Naive Bayes classifier via the EM algorithm with a class distribution constraint. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003); Edmonton, Canada*, pages 127–134. Morgan Kaufmann, 2003.

[146] H. van Deutekom, P. Supply, P.E.W. de Haas, E. Willery, S.P. Hoijng, C. Locht, R.A. Coutinho, and D. van Soolingen. Molecular typing of *Mycobacterium tuberculosis* by mycobacterial interspersed repetitive unit-variable-number tandem repeat analysis, a more accurate method for identifying epidemiological links between patients with tuberculosis. *J Clin Microbiol*, 43:4473–4479, 2005.

[147] J.D.A. van Embden, M.D. Cave, J.T. Crawford, J. Dale, K.D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, T.M. Shinnikl, and P.M. Small. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol*, 31:406–409, 1993.

[148] J.D.A. van Embden, T. van Gorkom, K. Kremer, R. Jansen, B.A. van Der Zeijst, and L.M. Schouls. Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J Bacteriol*, 9:2393–2401, 2000.

[149] D. van Soolingen, D.W. Hermans, P.E. de Haas, D.R. Soll, and J.D.A. van Embden. Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol*, 29:2578–2586, 2001.

[150] D. van Soolingen, L. Qian, P.E. Haas, J.T. Douglas, H. Traore, and F. Portaels et al. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. *J Clin Microbiol*, 33:3234–3238, 1995.

[151] A.M. Vos, A. Meima, S. Verver, C.W.N. Looman, V. Bos, M.W. Borgdorff, and J.D.F. Habbema. High incidence of pulmonary tuberculosis persists a decade after immigration, the Netherlands. *Emerg Infect Dis*, 10, 2004.

[152] E. Vynnycky, N. Nagelkerke, M.W. Borgdorff, D. van Soolingen, J.D.A. van Embden, and P.E. Fine. The effect of age and study duration on the relationship between "clustering" of DNA fingerprint patterns and the proportion of tuberculosis disease attributable to recent transmission. *Epidemiol Infect*, 126:43–62, 2001.

[153] R.M. Warren, S.L. Sampson, M. Richardson, G.D., van der Spuy, C.J. Lombard, T.C. Victor, and P.D. van Helden. Mapping of IS*6110* flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. *Mol Microbiol*, 37:1405–1416, 2002.

[154] R.M. Warren, E.M. Streicher, S.L. Sampson, G.D. van der Spuy, M. Richardson, D. Nguyen, M.A. Behr, T.C. Victor, and P.D. van Helden. Microevolution of the Direct Repeat region of *Mycobacterium tuberculosis*: Implications for interpretation of spoligotyping data. *J Clin Microbiol*, 40:4457–4465, 2002.

[155] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17:309–318, 2001.

[156] S. Zhong and J. Ghosh. A unified framework for model-based clustering. *Journal of Machine Learning Research*, 4:1001–1037, 2003.

[157] P.L. Zuber, M.T. McKenna, N.J. Binkin, I.M. Onorato, and K.G. Castro. Long-term risk of tuberculosis among foreign-born persons in the US. *JAMA*, 278:304–307, 1997.