PROBABILISTIC MODELS FOR PHYLOGENETIC CLASSIFICATION OF *MYCOBACTERIUM TUBERCULOSIS* COMPLEX GENOTYPING DATA

By

James Blondin

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COMPUTER SCIENCE

Approved:

____
Kristin P. Bennett, Thesis Adviser

____
Petros Drineas, Thesis Committee Member

____
Bülent Yener, Thesis Committee Member

Rensselaer Polytechnic Institute
Troy, New York

April 2013
(For Graduation May 2013)

CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank Nalin Rastogi and David Couvin at the Intstitut Pasteur de la Guadeloupe for supplying data as well as expertise in compiling the details of the lineage hierarchy. Also, thanks to Lauren Cowan of the United States Centers for Disease Control and Prevention for providing both data and expertise. This work was supported by NIH R01LM009731.

Thank you to all my peers and friends at RPI. Thank you as well to my professors and the exceedingly helpful staff within both the Computer Science and Mathematical Sciences departments.

Finally, I would like to extend my gratitude to my advisor Prof. Kristin P. Bennett, whose guidance in both this research project and in my other research and educational pursuits has always been extremely valuable and greatly appreciated.

ABSTRACT

This thesis presents a semi-supervised hierarchical Bayesian network to classify strains of *Mycobacterium Tuberculosis* complex (MTBC) into a three-tier set of genetic lineages and sublineages. MTBC is the causative agent of the infectious disease Tuberculosis (TB), which resulted in over 1.4 million deaths in 2011. Two main types of DNA fingerprinting techniques—spacer oligonucleotide typing (spoligotyping) and mycobacterial interspersed repetitive units (MIRUs)—are regularly used by public health officials and TB researchers to track and control TB.

The model and algorithms presented in this thesis use spoligotype and MIRU data combined from multiple heterogeneous data sources labeled by different experts to provide a model that is able to classify MTBC isolates into a hierarchical phylogenetic structure. The model is trained on over 117064 isolate DNA fingerprints collected by the United States Centers for Disease Control and Prevention, the SITVITWEB database at Institut Pasteur de Guadeloupe, and the MIRU-VNTR*plus* collection of MTBC strains. The model achieves high classification accuracy, confirming many well-established lineages at all hierarchy levels, and provides visualizations of spoligotype and MIRU signatures for each lineage. In addition, the model discovers some inconsistencies in MTBC labels between data sources, and suggests possible resolutions of these inconsistencies. After further study and refinement, this approach will form the basis for a new tool for MTBC lineage identification freely available online.

CHAPTER 1
Introduction

This chapter provides an overview of the research presented in this thesis and provides some background on tuberculosis epidemiology and existing efforts to model the strains of bacteria that cause tuberculosis.

1.1 Overview

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* complex (MTBC) bacteria. TB infects one-third of the world population resulting in over 1.4 million deaths per year. DNA fingerprinting of MTBC has become a routine part of TB control and surveillance. For example, the United States Centers for Disease Control and Prevention (CDC) calls for DNA fingerprinting of MTBC isolates from all culture-positive TB patients in the United States (US). Public health officials use these DNA fingerprints to help establish or eliminate potential TB transmission between individuals, to help identify TB outbreaks, and to track the evolution of TB on a population level. Thus, large databases of DNA fingerprints have been amassed that reflect the state of TB worldwide. This thesis works with a combination of databases gathered from the United States Centers for Disease Control and Prevention, the SITVITWEB project at Institute Pasteur de Guadeloupe (IP), and the MIRU-VNTR*plus* project [1, 2, 3, 4].

This study focuses on two main types of DNA fingerprinting globally used for MTBC genotyping and MTBC epidemiology: spacer oligonucleotide typing ("spoligotyping") [5] and mycobacterial interspersed repetitive units – variable-number-tandem-repeat (MIRU-VNTR, or simply MIRU) [6].

In molecular epidemiology, MTBC isolates are typically analyzed by phylogenetic lineages or clades. The lineages can be definitively determined by long sequence polymorphisms (LSPs) or regions of deletion (RDs), but these LSPs and RDs are not used for DNA fingerprinting since they do not have enough variance in a population [7, 8]. Thus, the primary goal of this work is to predict lineages using only spoligotypes and/or MIRU that are available. The lineages have a known hierarchical or multi-tier structure. The top-level or major lineages have been defined by LSPs and RDs and have been well established. Experts have used spoligotypes to determine sublineages and some of these sublineages have been validated by LSPs and RDS [9]. The definitions, granularity, and naming of sublineages can vary by organization: CDC uses a more coarse sublineage definition which we call mid-level lineages while IP uses a finer sublineage definition which we call sub-level classification.

The goal of this thesis is to create a model to predict all three levels of the lineage hierarchy. The lineage hierarchy is provided in Tables 1.1 and 1.2. Since the three data sets do not provide all three tiers of labels for each example, this work uses a semi-supervised method to handle the partially-labeled data. The model must handle isolates with spoligotype and MIRU as well as spoligotypes only, since much more data exists for the older spoligotypes. A further challenge occurs from the fact that the lineage labels maybe wrong since they typically are determined by experts using spoligotypes only [10]. The hope is that a semi-supervised spoligotype and MIRU model will correct lineage labels caused by convergent evolution of spoligotypes or expert error or disagreement.

This thesis builds upon previous unsupervised methods [11] for sub-lineage prediction using spoligotype only and supervised top-level lineage prediction using spoligotype and MIRU [12, 13].

1.2 MTBC Genotyping

We review the two DNA fingerprinting methods.

Spoligotyping is based on the changes in the direct repeat locus of the MTBC chromosome; the “spoligotype” of an MTBC isolate indicates the status of 43 possible polymorphisms in this locus. In the data recorded for an MTBC isolate, these polymorphisms are indicated as a simple Boolean flag indicating whether or not a particular oligonucleotide spacer exists in the DNA fingerprint of this isolate.

MIRU genotyping is also based on certain DNA properties of the MTBC chromosome. In the case of the MIRU measurements, the count of variable number tandem repeats (VNTRs) is tallied at various loci in the chromosome. In this thesis, we consider a set of 12 MIRU loci commonly used for MTBC DNA fingerprinting [6]. These 12 MIRU loci used here are labeled in the literature as: MIRU2, MIRU4, MIRU10, MIRU16, MIRU20, MIRU23, MIRU24, MIRU26, MIRU27, MIRU31, MIRU39 and MIRU40. Each of these MIRU loci is represented in data sets as a (usually) numeric value that indicates the number of VNTRs at that MIRU locus.

1.3 Discussion of Data Set

The data set used in this thesis consists of 15851 separate DNA fingerprints of MTBC found in culture-positive TB patients, each having a unique spoligotype-only or spoligotype-MIRU combination. As each of these data examples can occur in several different TB patients and thus have a cardinality greater than one, this full data set consists of 117064 isolates of MTBC.

Section 1.3.1 discusses the sources that were combined to create this data set as well as the the classification lineages used in these data sets. Each of the lineages are describing using up to three labels: a top level label (referred to in the data set as `Ctop`), and mid-level label (`Cmid`), and a sub-level label (`Csub`). The purpose of this three-level approach is to to organize the labels of strains of MTBC into a format which more closely matches existing phylogenetic trees used by various MTBC experts.

1.3.1 Data Set Sources and Lineages

The data set analyzed in this thesis comes from three sources. The Centers for Disease Control and Prevention (CDC) provide spoligotype and MIRU information for TB patients in the US. The Institut Pasteur de la Guadeloupe provides an extract from their SITVITWEB online data set [2], courtesy of Nalin Rastogi and David Couvin. The final data source is extracted from the MIRU-VNTR*plus* online database [3, 4].

Each data set source provides different labels. In order to combine these data sets into a single set of data with a three-tier lineage classification structure, the labels provided by the the different data sources were combined / inferred as necessary. The details of combining these data set labels are described in the following few paragraphs and result in the labels displayed in Tables 1.1 and 1.2. Table 1.1 displays those lineages that are typically referred to as the “modern” lineages, and Table 1.2 displays those lineages typically referred to as the “ancestral” lineages [14].

The labels provided with the CDC data set classified each isolate with a family and a subfamily. The family label was used to provide most of the `Ctop` data labels in Tables 1.1 and 1.2, and the subfamily label was used for the `Cmid` data labels. Since the CDC labels only provide a fairly coarse classification of the isolates, the `Csub` label for this data set was usually left as unlabeled.

The labels used in the Institut Pasteur SITVITWEB data set provide a fine-grained classification of each data point, with between 60 and 70 total classification labels. These classification labels provide the basis for the `Csub` data labels representing the sub-level labels in the data set. For the `Cmid` and `Ctop` were then inferred from these `Csub` data labels.

The labels used in the MIRU-VNTR*plus* data set provide a single lineage classification labels, which are generally equivalent to the `Csub` and `Cmid` labels found in the CDC and SITVITWEB data sets. In some cases, the MIRU-VNTR*plus* data set provided additional `Ctop` labels, particularly for some of the smaller lineages of MTBC such as *M. microti* and *M. pinnipedii.*

These labels include recent findings which specify the separation of a subset of the Euro-American lineage into a distinct mid-level lineage that contains several related strains from Africa: LAM10-CAM (Cameroon), S, and T2-uganda [9, 15]. Other recent publications describe some changes to the Haarlem mid-level lineage, which introduces the sub-level lineages Ural-1 and Ural-2 [16]. These label changes are included in Tables 1.1 and 1.2, and have been applied to the data from all three sources (where applicable).

1.3.2 Missing Data and Labels

The combined data set possesses certain characteristics which must be taken into consideration during model and training algorithm development. Specifically,

Table 1.1: Modern three-level MTBC strain lineages.

`Ctop`	`Cmid`	`Csub`
East-African Indian	East-African Indian	CAS1-Delhi
		CAS1-Kili
		CAS2
East Asian (Beijing)	East Asian (Beijing)	Beijing
Euro-American	EuroAm-African	LAM10-CAM
		S
		T2-uganda
	Haarlem	H1
		H2
		H3
		Ural-1
		Ural-2
	LAM	LAM1
		LAM11-ZWE
		LAM12-Madrid1
		LAM2
		LAM3
		LAM4
		LAM5
		LAM6
		LAM7-TUR
		LAM8
		LAM9
	T	H37Rv
		T1
		T1-RUS2
		T2
		T3
		T3-ETH
		T3-OSA
		T4
		T4-CEU1
		T5
		T5-Madrid2
		T5-RUS1
		T-tuscany
	X	X1
		X2
		X3

one or more of the data features (specifically in the MIRU data) may be missing, and parts or even all of the classification labels may be unknown for a particular isolate.

Of the two genotyping methods (spoligotyping and MIRU fingerprinting), spoligotyping was developed earlier; as a result, many MTBC genotype data sets have a limited number of isolates with both MIRU and spoligotype type data. The rest of the data only provides spoligotype information, leaving the MIRU values blank. In order to use all the data in the combined data set, any model and training algorithm must be able to properly handle this missing data.

Table 1.2: Ancestral three-level MTBC strain lineages.

`Ctop`	`Cmid`	`Csub`
Indo-Oceanic	Bangladesh	EAI6-BGD1 EAI7-BGD2
	India	EAI3-IND
	Manila	EAI2-Manila
	Mexico	EAI-Mexico
	Nonthaburi	EAI2-nonthaburi
	Vietnam	EAI4-VNM
	Unknown Mid-level	EAI1-SOM EAI2 EAI8-MDG
Mycobacterium africanum	West African 1	AFRI_2 AFRI_3
	West African 2	AFRI_1
Mycobacterium bovis	*Mycobacterium bovis*	BOV_1 BOV_2 BOV_3
Mycobacterium canettii	*Mycobacterium canettii*	Canettii
Mycobacterium caprae	*Mycobacterium caprae*	Caprae
Mycobacterium microti	*Mycobacterium microti*	Microti
Mycobacterium mungi	*Mycobacterium mungi*	M. mungi
Mycobacterium pinnipedii	*Mycobacterium pinnipedii*	Pini1 Pini2

Given the fact that the data originally comes from multiple data sources which do not have the same number of data labels as the final combined data set, many of the data labels are missing one or more of the `Ctop`, `Cmid`, and `CSub` data labels. For example, data from the CDC data set cannot typically be mapped down to the `Csub` level, meaning that while the `Ctop` and `Cmid` for a CDC isolate are set, the `Csub` remains unlabeled. Thus, a standard supervised learning algorithm for estimating the parameter of a model will be insufficient for this data set, requiring the use of semi-supervised estimation techniques.

1.4 Existing Tuberculosis Lineage Models

Several models have already been developed to either classify or cluster MTBC data into lineages. The SPOTCLUST approach [11] describes an unsupervised model and learning algorithm to determine sublineage labels of MTBC strains. SPOTCLUST models the MTBC spoligotype data as a mixture model, and uses the widely-used expectation-maximization (EM) algorithm to estimate the model's parameters. This method provides the basis for much of the work in this thesis.

A fully-supervised approach using a conformal Bayesian network (CBN) is presented in [12]. This approach uses both spoligotype and MIRU data to classify the top-level lineages for MTBC strains. Another supervised approach presented in [13] provides a set of rules for classifying the top-level lineages, also using both spoligotype and MIRU data.

1.5 Semi-supervised Approach for MTBC Lineage Modeling

This thesis builds upon the approaches described in Section 1.4 by extending the single-level models into a multi-tier model which can capture the top-level, mid-level, and sub-level lineages available in the combined data set. The goal of this work is to utilize all of the spoligotype and MIRU data available to provide a model that effectively classifies new MTBC strains with all three levels of labels, providing a basic phylogenetic tree for each strain.

In order to accommodate the specific characteristics of the data set described in Section 1.3, and in particular the fact that much of the data set is partially-labeled (as described in Section 1.3.2), the approach described in this thesis uses a semi-supervised parameter estimation algorithm based on the EM algorithm commonly used in unsupervised data clustering tasks.

1.6 Structure

Chapter 2 describes the three-tier lineage model, building upon a simpler single-tier model similar to that described in [11]. Chapter 3 discusses the EM algorithm used to estimate the parameters in the models described in Chapter 2. Chapter 4 show the results form comparing the effectiveness of the single-tier and multi-tier models on the data set described in Section 1.3, following by a conclusion and references.

Chapter A in the appendix includes details that supplement Chapter 3 by providing extra mathematical derivations of the EM update steps.

CHAPTER 2
Hierarchical Model for MTBC Classification

This chapter reviews potential models for phylogenetic classification and clustering of MTBC strains using hierarchical Bayesian models. Algorithms used to estimate the parameters of these models are described in Section 3.1.

2.1 Overview

Two styles of models for the modeling of MTBC strains are discussed in this thesis: a single-tier model that provides a one-level phylogenetic classification; and a hierarchical three-tier model that provides a top-level lineage, a mid-level lineage, and a sub-level lineage for each isolate.

The single-tier model is based on the SPOTCLUST algorithm [11], and provides a model allowing for a single-label classification / clustering of MTBC strains based on their spoligotypes and MIRU values. The basic set of labels considered for this model are only those listed in the `Csub` column in Table 1.1 and Table 1.2, ignoring the `Cmid` and `Ctop` labels. The single-tier model in this thesis uses a similar approach to SPOTCLUST, expanding beyond that work by including MIRU values. Additionally, as discussed in Section 3.1.4, the training of these models incorporates existing data labels in a semi-supervised fashion.

The multi-tier model extends this single-tier model by providing a three-tier phylogenetic structure for the classification / clustering of MTBC strains. For the purposes of this model, each isolate in the data set can have up to three labels: a top-level lineage, and mid-level lineage, and a sub-level lineage, with the possibility that one or more of these labels is unlabeled. The base set of labels used in this models is provided in Tables 1.1 and 1.2.

Section 2.2 discusses general notation used in all the described models. Section 2.3 describes a simple single-tier model where the characteristics of a strain depend only upon a single label. Section 2.4 describes a slightly more complicated single-tier model that incorporates a “hidden parent”, which utilizes some knowledge of how spoligotypes evolve in strains of MTBC. Section 2.5 describes the multi-tier model.

2.2 General Notation

Let n be the number of data examples in a data set, where each example $\mathbf{x}_i = (\mathbf{s}_i, \mathbf{m}_i)$ for $i = 1, \ldots, n$ is composed of spoligotype measurements $\mathbf{s}_i$ and MIRU measurements $\mathbf{m}_i$. Let X be the entire set of n data examples, $X = (S, M)$, where S is the set of all spoligotype measurements and M is the set of all MIRU measurements.

Each spoligotype measurement is a set of D_S binary values, $\mathbf{s}_i \in \{0, 1\}^{D_S}$, and each MIRU measurement is a set of D_M categorical values, $\mathbf{m}_i \in \{q_1, \ldots, q_R\}^{D_M}$. In this case, D_S is the number of spoligotype values for a single example, D_M is the number of MIRU values, and R is the number of different categories the MIRU value can take.

For our data, $R = 20$, the twenty different categories a MIRU value can take: $\{0, \ldots, 9, A, R, S, T, U, V, W, X, Y, Z\}$. These represent the MIRU values 0-9, 10+ (represented by A), and R-Z. Additionally, for our data, we have 43 spoligotype measurements and 12 MIRU values, so $D_S = 43$ and $D_M = 12$.

We use the Iverson bracket notation $[P]$, which specifies a function that returns 1 when the logical statement P within the square brackets is true, and 0 otherwise.

2.3 Single-Tier Model

The simplest model that we consider is a single-level probabilistic classification model similar to a naïve Bayes classifier or mixture model, wherein the probability of each spoligotype or MIRU measurement is solely dependent on the label of the data example, and independent of the other measurements.

Let C be a random variable representing the classification of an isolate, distributed as a categorical variable with k possible classifications $\{c_1, \ldots, c_k\}$.

The probability of a particular data example with measurements $\mathbf{s}_i, \mathbf{m}_i$ given

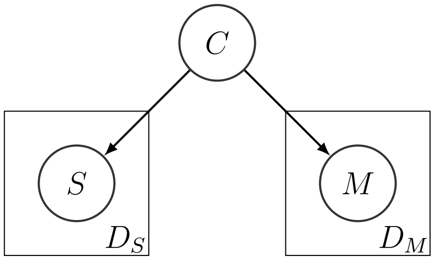


Figure 2.1: Template Diagram for Single-Tier Sublineage Model.

a set of probability parameters $\mathbf{\Theta}$, summed over the k possible classifications, is

$$p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) = \sum_{j=1}^{k} p(c_j) p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) \ .$$

The set of probability parameters $\mathbf{\Theta}$ is composed of the probability parameters which describe the component conditional distributions. The probability $p(\mathbf{s}_i \mid c_j)$ is a set of D_S Bernoulli distributions given the class, one for each of the spoligotype spacer, where each individual distribution given the class is $p(s_{id} \mid c_j)$. This results in the probability

$$p(\mathbf{s}_i \mid c_j) = \prod_{d=1}^{D_S} p(s_{id} \mid c_j) = \prod_{d=1}^{D_S} \sigma_{jd}^{s_{id}} (1 - \sigma_{jd})^{1-s_{id}}$$

where σ_{jd} is the probability of existence of the dth spoligotype spacer given class c_j.

The probability $p(\mathbf{m}_i \mid c_j)$ is a set of D_M categorical distributions (also called discrete, multinoulli, or sometimes multinomial distributions) given the class, resulting in the probability

$$p(\mathbf{m}_i \mid c_j) = \prod_{d=1}^{D_M} p(m_{id} \mid c_j) = \prod_{d=1}^{D_M} \mu_{jd1}^{[m_{id}=q_1]} \cdots \mu_{jdR}^{[m_{id}=q_R]}$$

wherein μ_{jdr} is the probability of the dth MIRU locus having value q_r (the rth category) given class c_j.

Finally, the probability $p(c_j)$ is the is simply the mixture weight $p(c_j) = \alpha_j$.

Combined, these parameters make up the parameter set

$$\mathbf{\Theta} = \{\alpha_j\}_{j=1,\ldots,k} \cup \{\sigma_{jd}\}_{j=1,\ldots,k,\, d=1,\ldots,D_S} \cup \{\mu_{jdr}\}_{j=1,\ldots,k,\, d=1,\ldots,D_M,\, r=1,\ldots,R} \,.$$

The log-likelihood of the parameters $\mathbf{\Theta}$ is

$$\begin{aligned}\log L(\mathbf{\Theta}) = \log\{p(S, M \mid \mathbf{\Theta})\} &= \log\left\{\prod_{i=1}^{n} p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right\} = \sum_{i=1}^{n} \log\left\{p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right\} \\ &= \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{k} p(c_j) p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j)\right\} \\ &= \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{k} \alpha_j \prod_{d=1}^{D_S} \sigma_{jd}^{s_{id}} (1 - \sigma_{jd})^{1 - s_{id}} \prod_{d=1}^{D_M} \mu_{jd1}^{[m_{id}=q_1]} \cdots \mu_{jdR}^{[m_{id}=q_R]}\right\} \,.\end{aligned}$$

Maximization of this log-likelihood is difficult due to the summation over the k possible label classifications. However, in our semi-supervised scenario, some of these labels are known and the probability $p(c_j)$ is fixed at a value of 1 or 0 for those data examples. For the unlabeled data points, we can include latent variables specifying the choices of label for each of the unclassified data examples and maximize over the expectation of these latent variables using a standard expectation-maximization algorithm. This is described in detail in Section 3.1.1.

2.4 Single-Tier Sublineage Model With Hidden Parent Assumption

In existing studies of spoligotype evolution among strains of MTBC it has been determined that the evolution of the strains typically results in the deletion of one or more spacers. The insertion of new spacers as a very rare event. [17, 18, 19, 20]. The SPOTCLUST algorithm [11] introduced the concept of "Hidden Parents" to model this phenomenon.

The idea behind the Hidden Parent is that if the distribution of a particular class c_j specifies that a particular spoligotype spacer is present ($s_{id} = 1$) with high probability, we should allow an observed isolate of that class to drop that spacer with some probability greater than 0. However, if the distribution of a class c_j

specifies that a particular spoligotype spacer is absent, the chance that an observed isolate of class c_j has that particular spacer present should be very low.

To do this, we introduce a new hidden (and predefined) random variable between the class layer and the spoligotype layer, as seen in Figure 2.2.

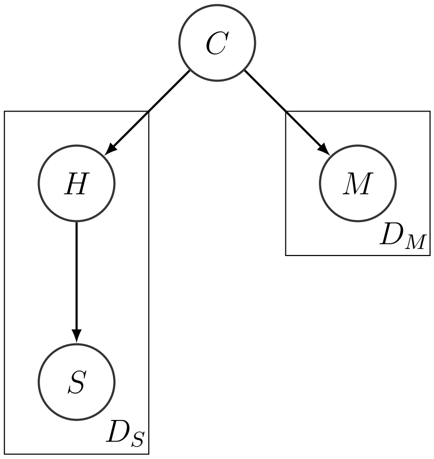


Figure 2.2: Template Diagram for Single-Tier Model with Hidden Parent Assumption.

With this hidden parent, the likelihood of a measurement i is

$$p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) = \sum_{j=1}^{k} p(c_j)p(\mathbf{s}_i \mid c_j)p(\mathbf{m}_i \mid c_j) = \sum_{j=1}^{k} p(c_j)p(\mathbf{s}_i \mid \mathbf{h}_i)p(\mathbf{h}_i \mid c_j)p(\mathbf{m}_i \mid c_j)$$

where k is the number of possible values of C (classifications), and $\mathbf{\Theta}$ is the set of probability parameters

$$\mathbf{\Theta} = \{\alpha_j\}_{j=1,\ldots,k} \cup \{\sigma_{jd}\}_{j=1,\ldots,k,\, d=1,\ldots,D_S} \cup \{\mu_{jdr}\}_{j=1,\ldots,k,\, d=1,\ldots,D_M,\, r=1,\ldots,R}$$

with $p(c_j) = \alpha_j$,

$$p(\mathbf{h}_i \mid c_j) = \prod_{d=1}^{D_S} \sigma_{jd}^{h_{id}} (1 - \sigma_{jd})^{1-h_{id}},$$

and

$$p(\mathbf{m}_i \mid c_j) = \prod_{d=1}^{D_M} \mu_{jd1}^{[m_{id}=q_1]} \cdots \mu_{jdR}^{[m_{id}=q_R]} \ .$$

The Hidden Parent is established with the predefined parameters

$$\begin{aligned} p(s_{id} = 1 \mid h_{id} = 1) &= \eta_{11} \ , & p(s_{id} = 0 \mid h_{id} = 0) &= \eta_{00} \ , \\ p(s_{id} = 0 \mid h_{id} = 1) &= \eta_{01} = 1 - \eta_{11} \ , & p(s_{id} = 1 \mid h_{id} = 0) &= \eta_{10} = 1 - \eta 00 \end{aligned}$$

where we set $\eta_{11} = 0.9$ and $\eta_{10} = 10^{-7}$. These choices for η_{11} and $\eta 10$, as suggested in [11], enforce the asymmetric reliance of a spoligotype spacer upon the class: having $p(s_{id} = 1 \mid h_{id} = 1) = 0.9$ allows for some chance of spoligotypes losing a spacer even if $p(h_{id} \mid c_j) = \sigma_{jd}$ is high. Conversely, having $p(s_{id} = 1 \mid h_{id} = 0) = 10^{-7}$ prevents a spoligotype having a spacer if $p(h_{id} \mid c_j) = \sigma_{jd}$ is low.

We can rewrite $p(s_{id} \mid h_{id})$ as

$$p(s_{id} \mid h_{id}) = (\eta_{11} h_{id} + \eta_{10}(1 - h_{id}))^{s_{id}} (\eta_{01} h_{id} + \eta_{00}(1 - h_{id}))^{1 - s_{id}}$$

which allows us to compute $p(\mathbf{s}_i \mid c_j)$ as

$$\begin{aligned} p(\mathbf{s}_i \mid c_j) &= \prod_{d=1}^{D_S} \sum_{\hat{h}_{id} \in \{0,1\}} p(s_{id} \mid h_{id} = \hat{h}_{id}) p(h_{id} = \hat{h}_{id} \mid c_j) \\ &= \prod_{d=1}^{D_S} \left(\eta_{11}^{s_{id}} \eta_{01}^{1 - s_{id}} \sigma_{jd} + \eta_{10}^{s_{id}} \eta_{00}^{1 - s_{id}} (1 - \sigma_{jd}) \right) \\ &= \prod_{d=1}^{D_S} (\eta_{11} \sigma_{jd} + \eta_{10}(1 - \sigma_{jd}))^{s_{id}} (\eta_{01} \sigma_{jd} + \eta_{00}(1 - \sigma_{jd}))^{1 - s_{id}} \ . \end{aligned}$$

The log-likelihood of the parameters $\mathbf{\Theta}$ is

$$\begin{aligned}
\log L(\mathbf{\Theta}) &= \log \left\{ p(S, M \mid \mathbf{\Theta}) \right\} \\
&= \log \left\{ \prod_{i=1}^{n} p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} \\
&= \sum_{i=1}^{n} \log \left\{ p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} \\
&= \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} p(c_j) p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) \right\} \\
&= \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \alpha_j \prod_{d=1}^{D_S} \left(\eta_{11}^{s_{id}} \eta_{01}^{1-s_{id}} \sigma_{jd} + \eta_{10}^{s_{id}} \eta_{00}^{1-s_{id}} (1 - \sigma_{jd}) \right) \right. \\
&\qquad\qquad \left. \prod_{d=1}^{D_M} \mu_{jd1}^{[m_{id}=q_1]} \cdots \mu_{jdR}^{[m_{id}=q_R]} \right\} .
\end{aligned}$$

Details of the expectation-maximization algorithm used to estimate these parameters are found in Section 3.1.2.

2.5 Multi-Tier Model With Hidden Parent Assumption

The other style of model we are considering is a hierarchical three-tier model. This model is intended to correspond to the three-level data set labels described in Section 1.3. Each isolate has three labels specifying its top-level, mid-level,and sub-level classification.

Let A be the top-level classification of the measurement with k_A possible classifications $a_1, \ldots, a_{k_A}$. Let B be the mid-level classification of a measurement with k_B possible classifications $b_1, \ldots, b_{k_B}$. Let C be the sub-level classification of a measurement with k_C possible classifications $c_1, \ldots, c_{k_C}$.

The probability of a measurement i given the model diagrammed in Figure 2.3 is

$$p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) = \sum_{u=1}^{k_A} \sum_{v=1}^{k_B} \sum_{j=1}^{k_C} p(a_u) p(b_v \mid a_u) p(c_j \mid b_v) p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j)$$

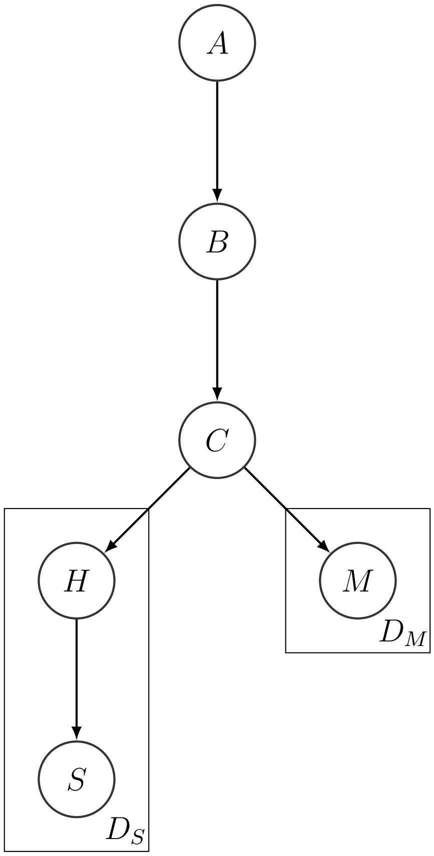


Figure 2.3: Template Diagram for Multi-Tier Model with Hidden Parent Assumption.

where $\mathbf{\Theta}$ is the set of probability parameters

$$\begin{aligned}\mathbf{\Theta} = \{\alpha_u\}_{u=1,\ldots,k_A} \cup \{\beta_{vu}\}_{v=1,\ldots,k_B,\, u=1,\ldots,k_A} \cup \{\gamma_{jv}\}_{j=1,\ldots,k_C,\, v=1,\ldots,k_B} \\ \cup \{\sigma_{jd}\}_{j=1,\ldots,k_C,\, d=1,\ldots,D_S} \cup \{\mu_{jdr}\}_{j=1,\ldots,k_C,\, d=1,\ldots,D_M,\, r=1,\ldots,R}\end{aligned}$$

with $p(a_u) = \alpha_u$, $p(b_v \mid a_u) = \beta_{vu}$, $p(c_j \mid b_v) = \gamma_{jv}$,

$$p(\mathbf{s}_i \mid c_j) = \prod_{d=1}^{D_S} \left(\eta_{11}\sigma_{jd} + \eta_{10}(1-\sigma_{jd})\right)^{s_{id}} \left(\eta_{01}\sigma_{jd} + \eta_{00}(1-\sigma_{jd})\right)^{1-s_{id}}$$

as computed in Section 2.4, and

$$p(\mathbf{m}_i \mid c_j) = \prod_{d=1}^{D_M} \mu_{jd1}^{[m_{id}=q_1]} \cdots \mu_{jdR}^{[m_{id}=q_R]} \ .$$

The predefined parameters η_{ij} are used to govern the hidden parent assumption, and are defined in Section 2.4 as

$$\begin{aligned} &p(\mathbf{s}_{id} = 1 \mid \mathbf{h}_{id} = 1) = \eta_{11} \ , && p(\mathbf{s}_{id} = 0 \mid \mathbf{h}_{id} = 0) = \eta_{00} \ , \\ &p(\mathbf{s}_{id} = 0 \mid \mathbf{h}_{id} = 1) = \eta_{01} = 1 - \eta_{11} \ , && p(\mathbf{s}_{id} = 1 \mid \mathbf{h}_{id} = 0) = \eta_{10} = 1 - \eta 00 \ . \end{aligned}$$

We again set $\eta_{11} = 0.9$ and $\eta_{10} = 10^{-7}$.

The log-likelihood of the parameters $\mathbf{\Theta}$ is

$$\begin{aligned} \log L(\mathbf{\Theta}) &= \log \left\{ p(S, M \mid \mathbf{\Theta}) \right\} \\ &= \log \left\{ \prod_{i=1}^{n} p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} \\ &= \sum_{i=1}^{n} \log \left\{ p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} \\ &= \sum_{i=1}^{n} \log \left\{ \sum_{u=1}^{k_A} \sum_{v=1}^{k_B} \sum_{j=1}^{k_C} p(a_u) p(b_v \mid a_u) p(c_j \mid b_v) p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) \right\} \\ &= \sum_{i=1}^{n} \log \left\{ \sum_{u=1}^{k_A} \alpha_u \sum_{v=1}^{k_B} \beta_{vu} \sum_{j=1}^{k_C} \gamma_{jv} p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) \right\} \end{aligned}$$

Details of the expectation-maximization algorithm used to estimate these parameters are found in Section 3.1.3.

An important note to consider is that the model described in this section allows for more freedom than the lineage hierarchy displayed in Tables 1.1 and 1.2. In this model, each isolate could conceivably be classified with any one of the possible top-level, mid-level and sub-level labels. As a result, this model could classify an isolate as having a `CSub` or `Cmid` label that does not correspond to the `Ctop` label as expected by the hierarchy. As we discuss in Section 4.2, this has some interesting consequences.

CHAPTER 3
Methodology

This chapter discusses the methods used to estimate the parameters of the models described in Chapter 2. Section 3.1 describes the details of the expectation-maximization (EM) algorithm as applied to the single-tier and multi-tier models. Section 3.2 describes the methods used to initialize the EM algorithm (to avoid getting stuck in local maxima) along with the techniques used to validate the model. Section 3.3 describes the techniques used to select the appropriate number of available extra top-level, mid-level, and sub-level labels.

3.1 Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm is an iterative method for solving difficult maximum-likelihood problems [21, 22]. The general EM approach presupposes that for each data example without a label, there exists an unobserved data label (called a *latent variable*). By doing so, the calculation of the maximization of the likelihood becomes tractable.

In order to compute the maximization of the likelihood, we must first compute the expected value of the latent variables given a certain set of observed (non-latent) variables as well as an existing parameterization of the underlying model. Note that for the first iteration of this algorithm, the model parameterization must be set to some initial value.

Once the computation of the expectation of the latent variables is complete, the likelihood maximization of the model parameters given these expected values becomes easily computable.

Sections 3.1.1, 3.1.2, and 3.1.3 describe the computation of the expectation of the latent variables and the estimated parameter values that result from the maximization process. Section 3.1.4 describes the method used to extend the EM algorithm to be able to use partially-labeled data. Section 3.1.5 describes extensions to the EM update steps which allow for repeated isolates in the data set.

Section 3.1.6 describes the method used for dealing with missing MIRU values. Finally Section 3.1.7 discusses enhancements made to the EM algorithm which leverage prior distributions of the model parameters.

3.1.1 EM for Single-Tier Sublineage Model

Recall from Section 2.3 the observed log-likelihood of the parameters $\mathbf{\Theta}$ for a single-tier sublineage model,

$$\begin{aligned}\log L(\mathbf{\Theta}) &= \log\left\{\prod_{i=1}^{n} p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right\} \\ &= \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{k} \alpha_j \prod_{d=1}^{D_S} \sigma_{jd}^{s_{id}} (1-\sigma_{jd})^{1-s_{id}} \prod_{d=1}^{D_M} \mu_{jd1}^{[m_{id}=q_1]} \cdots \mu_{jdR}^{[m_{id}=q_R]}\right\} .\end{aligned}$$

The approach used for the derivation of the EM algorithm in this and the following sections is based on the method presented in [21].

Adding in a set of unobserved measurements Z which specifies the unobserved data (the choice c_j for each unlabeled measurement), we have a complete log-likelihood

$$\log L_C(\mathbf{\Theta}) = \log p(S, M, Z \mid \mathbf{\Theta}) .$$

This complete log-likelihood has an expectation, given the observed data and parameter estimates at a certain iteration t, expressed as $E_Z[\log L_C(\mathbf{\Theta}) \mid S, M, \mathbf{\Theta}^{(t)}]$. Thus, we seek to iteratively find the parameters that maximize this expectation

$$\mathbf{\Theta}^{(t+1)} = \arg\max_{\mathbf{\Theta}} \; E_Z[\log L_C(\mathbf{\Theta}) \mid S, M, \mathbf{\Theta}^{(t)}] .$$

The unobserved variables Z can be represented by a matrix $\mathbf{Z} \in \mathbb{R}^{n\times k}$ where z_{ij} is 1 if the data example i was generated by component j, and 0 otherwise.

To compute the complete log-likelihood, we find

$$\begin{aligned}\log L_C(\boldsymbol{\Theta}) = \log\{p(S, M, \mathbf{Z} \mid \boldsymbol{\Theta})\} &= \log\left\{\prod_{i=1}^{n}\left[\sum_{j=1}^{k} z_{ij} p(\mathbf{s}_i, \mathbf{m}_i \mid \boldsymbol{\Theta})\right]\right\} \\ &= \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{k} z_{ij} p(\mathbf{s}_i, \mathbf{m}_i \mid \boldsymbol{\Theta})\right\} \ .\end{aligned}$$

Because the latent variable $z_{ij} = 1$ for only one j, we can move the z_{ij} term and the inner summation out of the logarithm as follows:

$$\log L_C(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \log\{p(\mathbf{s}_i, \mathbf{m}_i \mid \boldsymbol{\Theta})\} \ .$$

Now we can compute

$$\begin{aligned}\boldsymbol{\Theta}^{(t+1)} &= \arg\max_{\boldsymbol{\Theta}} \ E_Z[\log L_C(\boldsymbol{\Theta}) \mid S, M, \boldsymbol{\Theta}^{(t)}] \\ &= \arg\max_{\boldsymbol{\Theta}} \ E_Z\left[\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \log\{p(\mathbf{s}_i, \mathbf{m}_i \mid \boldsymbol{\Theta})\} \,\middle|\, S, M, \boldsymbol{\Theta}^{(t)}\right] \\ &= \arg\max_{\boldsymbol{\Theta}} \ \sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \,\middle|\, \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \log\{p(\mathbf{s}_i, \mathbf{m}_i \mid \boldsymbol{\Theta})\} \ . \qquad (3.1)\end{aligned}$$

Equation 3.1 expresses the maximization problem which can be iteratively solved to find a local maximum parameter estimate for the single-tier model.

It is important to note that the expectation of z_{ij} can be computed as

$$\begin{aligned}E_Z\left[z_{ij} \,\middle|\, \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] &= 0 \cdot p(z_{ij} = 0 \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}) + 1 \cdot p(z_{ij} = 1 \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}) \\ &= p(z_{ij} = 1 \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}) \ ,\end{aligned}$$

which is the probability that measurement i is is labeled with classification j. This is equivalently

$$E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] = p(c_j \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)})$$

where

$$p(c_j \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}) = \frac{p(\mathbf{s}_i, \mathbf{m}_i \mid c_j)p(c_j)}{p(\mathbf{s}_i, \mathbf{m}_i)} = \frac{p(\mathbf{s}_i \mid c_j)p(\mathbf{m}_i \mid c_j)p(c_j)}{p(\mathbf{s}_i)p(\mathbf{m}_i)}$$

Now, we want to separately maximize for the parameters that make up the probability $p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})$: $\alpha_{\hat{j}}$ (for $\hat{j} = 1, \ldots, k$), $\sigma_{\hat{j}\hat{d}}$ (for $\hat{j} = 1, \ldots, k$ and $\hat{d} = 1, \ldots, D_S$), and $\mu_{\hat{j}\hat{d}\hat{r}}$ (for $\hat{j} = 1, \ldots, k$, $\hat{d} = 1, \ldots, D_M$, and $\hat{r} = 1, \ldots, R$). The derivations of the solutions of these maximization problems are listed in Section A.1.

The final parameter estimate updates are

$$\begin{aligned}
\alpha_{\hat{j}}^{(t+1)} &= \frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{n} , \\
\sigma_{\hat{j}\hat{d}}^{(t+1)} &= \frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] s_{i\hat{d}}}{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]} , \\
\mu_{\hat{j}\hat{d}\hat{r}}^{(t+1)} &= \frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{\sum_{r=1}^{R} \sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] [m_{i\hat{d}} = q_r]} .
\end{aligned}$$

The expectation-maximization algorithm begins by establishing an initial set of values $\mathbf{\Theta}^{(0)}$. Then, each of the expectations $E_Z[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}]$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ is computed and a new set of parameters $\mathbf{\Theta}^{t+1}$ is produced given the update steps above. These iterations continue until convergence is reached.

3.1.2 EM for Single-Tier Sublineage Model with Hidden Parent

From Section 2.4, the observed log-likelihood of the parameters $\mathbf{\Theta}$ is

$$\begin{aligned}
\log L(\mathbf{\Theta}) &= \log \left\{ \prod_{i=1}^{n} p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} \\
&= \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \alpha_j \prod_{d=1}^{D_S} \left(\eta_{11}^{s_{id}} \eta_{01}^{1-s_{id}} \sigma_{jd} + \eta_{10}^{s_{id}} \eta_{00}^{1-s_{id}} (1 - \sigma_{jd}) \right) \right. \\
&\qquad \left. \prod_{d=1}^{D_M} \mu_{jd1}^{[m_{id}=q_1]} \cdots \mu_{jdR}^{[m_{id}=q_R]} \right\} .
\end{aligned}$$

Similar to as in Section 3.1.1, if we add in a set of unobserved measures Z which specify the unobserved data (again, the choice c_j for each measurement), we have a complete log-likelihood

$$\log L_C(\mathbf{\Theta}) = \log p(S, M, Z \mid \mathbf{\Theta})$$

and we seek to find the parameters that maximize the expectation

$$\mathbf{\Theta}^{(t+1)} = \arg\max_{\mathbf{\Theta}} \; E_Z[\log L_C(\mathbf{\Theta}) \mid S, M, \mathbf{\Theta}^{(t)}] \, .$$

We again represent the unobserved variables as a matrix $\mathbf{Z} \in \mathbb{R}^{n \times k}$ where z_{ij} is 1 if the data example was generated by component j, and 0 otherwise. The expectation-maximization equation to solve is, as before,

$$\mathbf{\Theta}^{(t+1)} = \arg\max_{\mathbf{\Theta}} \sum_{i=1}^{n} \sum_{j=1}^{k} E_Z \left[z_{ij} \;\mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] \log \left\{ p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} \tag{3.2}$$

with

$$E_Z \left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] = p(c_j \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}) = \frac{p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) p(c_j)}{p(\mathbf{s}_i) p(\mathbf{m}_i)} \, .$$

Now, we want to separately maximize for the parameters that make up the probability $p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})$: $\alpha_{\hat{j}}$ (for $\hat{j} = 1, \ldots, k$), $\sigma_{\hat{j}\hat{d}}$ (for $\hat{j} = 1, \ldots, k$ and $\hat{d} = 1, \ldots, D_S$), and $\mu_{\hat{j}\hat{d}\hat{r}}$ (for $\hat{j} = 1, \ldots, k$, $\hat{d} = 1, \ldots, D_M$, and $\hat{r} = 1, \ldots, R$).

Section A.2 demonstrates the computation of the maximization these parameters that make up the probability $p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})$. The final parameter estimate updates are

$$\begin{aligned}
\alpha_{\hat{j}}^{(t+1)} &= \frac{\sum_{i=1}^{n} E_Z \left[z_{i\hat{j}} \;\mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right]}{n} \, , \\
\sigma_{\hat{j}\hat{d}}^{(t+1)} &= \frac{\sum_{i=1}^{n} E_Z \left[z_{i\hat{j}} \;\mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] (s_{i\hat{d}} - \eta_{10})}{(\eta_{11} - \eta_{10}) \sum_{i=1}^{n} E_Z \left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right]} \, , \\
\mu_{\hat{j}\hat{d}\hat{r}}^{(t+1)} &= \frac{\sum_{i=1}^{n} E_Z \left[z_{i\hat{j}} \;\mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right]}{\sum_{r=1}^{R} \sum_{i=1}^{n} E_Z \left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] [m_{i\hat{d}} = q_r]} \, .
\end{aligned}$$

As in Section 3.1.1, the expectation-maximization algorithm proceeds by first establishing an initial set of values $\mathbf{\Theta}^{(0)}$. Then, we iteratively compute each of the expectations $E_Z[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}]$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$, and produce a new set of parameters $\mathbf{\Theta}^{t+1}$ given these update steps. These iterations continue until convergence is reached.

3.1.3 EM for Multi-Tier Sublineage Model with Hidden Parent

Recall the three-tier model from Section 2.5, which has the observed log-likelihood of the parameters $\mathbf{\Theta}$

$$\begin{aligned}\log L(\mathbf{\Theta}) &= \log \left\{ \prod_{i=1}^{n} p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} \\ &= \sum_{i=1}^{n} \log \left\{ \sum_{u=1}^{k_A} \sum_{v=1}^{k_B} \sum_{j=1}^{k_C} p(a_u) p(b_v \mid a_u) p(c_j \mid b_v) p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) \right\} .\end{aligned}$$

Using a similar approach to the single-tier expectation-maximization process, we add a set of unobserved data measurements $W = X \cup Y \cup Z$ which specify the choices a_u,b_v, and c_j for each measurement. This gives the complete log-likelihood

$$\log L_C(\mathbf{\Theta}) = \log p(S, M, W \mid \mathbf{\Theta})$$

which has the expectation given the observed data $E_W[\log L_C(\mathbf{\Theta}) \mid S, M, \mathbf{\Theta}^{(t)}]$. We seek to find the parameters that maximize this expectation

$$\mathbf{\Theta}^{(t+1)} = \arg\max_{\mathbf{\Theta}} \; E_W[\log L_C(\mathbf{\Theta}) \mid S, M, \mathbf{\Theta}^{(t)}] \, .$$

The unobserved variables can be represented by matrices $\mathbf{X} \in \mathbb{R}^{n \times k_A}$, $\mathbf{Y} \in \mathbb{R}^{n \times k_B}$, and $\mathbf{Z} \in \mathbb{R}^{n \times k_C}$. Each value in the matrix (for example, x_{iu}) is 1 if the data example was labeled with that top-level, mid-level, or sub-level label, and 0 otherwise.

To compute the complete log likelihood, we find

$$\begin{aligned}\log L_C(\mathbf{\Theta}) &= \log\left\{p(S, M, W \mid \mathbf{\Theta})\right\} \\ &= \log\left\{\prod_{i=1}^{n}\left(\sum_{u=1}^{k_A} x_{iu}\sum_{v=1}^{k_B} y_{iv}\sum_{j=1}^{k_C} z_{ij}\, p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right)\right\} \\ &= \sum_{i=1}^{n}\log\left\{\sum_{u=1}^{k_A} x_{iu}\sum_{v=1}^{k_B} y_{iv}\sum_{j=1}^{k_C} z_{ij}\, p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right\}\ .\end{aligned}$$

Because $x_{iu} = 1$ for only one u, $y_{iv} = 1$ for only one v, and $x_{ij} = 1$ for only one j, we can move these terms and the inner summations out of the logarithm as follows:

$$\log L_C(\mathbf{\Theta}) = \sum_{i=1}^{n}\sum_{u=1}^{k_A} x_{iu}\sum_{v=1}^{k_B} y_{iv}\sum_{j=1}^{k_C} z_{ij}\log\left\{p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right\}\ .$$

Next, we can compute the maximization of the parameters given the expectation, written as

$$\begin{aligned}\mathbf{\Theta}^{(t+1)} &= \arg\max_{\mathbf{\Theta}}\ E_W[\log L_C(\mathbf{\Theta}) \mid S, M, \mathbf{\Theta}^{(t)}] \\ &= \arg\max_{\mathbf{\Theta}}\ E_W\left[\sum_{i=1}^{n}\sum_{\substack{u=1,\ldots,k_A\\ v=1,\ldots,k_B\\ j=1,\ldots,k_C}} x_{iu}y_{iv}z_{ij}\log\left\{p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right\}\,\middle|\, S, M, \mathbf{\Theta}^{(t)}\right] \\ &= \arg\max_{\mathbf{\Theta}}\ \sum_{i=1}^{n}\sum_{\substack{u=1,\ldots,k_A\\ v=1,\ldots,k_B\\ j=1,\ldots,k_C}} E_W\left[x_{iu}y_{iv}z_{ij}\ \middle|\ \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]\log\left\{p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right\} \quad (3.3)\end{aligned}$$

To compute the expectation of $x_{iu}y_{iv}z_{ij}$, we note that, much like in Section 3.1.1, the expectation of a particular set of labels is equivalent to the probability of those labels:

$$\begin{aligned} E_W\left[x_{iu}y_{iv}z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] &= p(x_{iu}=1, y_{iv}=1, z_{ij}=1 \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}) \\ &= p(a_u, b_v, c_j \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}) \\ &= \frac{p(a_u, b_v, c_j, \mathbf{s}_i, \mathbf{m}_i)}{p(\mathbf{s}_i, \mathbf{m}_i)} \\ &= \frac{p(\mathbf{s}_i \mid c_j)p(\mathbf{m}_i \mid c_j)p(c_j \mid b_v)p(b_v \mid a_u)p(a_u)}{p(\mathbf{s}_i)p(\mathbf{m}_i)} . \end{aligned}$$

Equation 3.3 is solved for an established $E_W\left[x_{iu}y_{iv}z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]$ by separate maximizing the parameters $\alpha_{\hat{u}}$, $\beta_{\hat{v}\hat{u}}$, $\gamma_{\hat{j}\hat{v}}$, $\sigma_{\hat{j}\hat{d}}$,and $\mu_{\hat{j}\hat{d}\hat{r}}$ separately. The derivations of these maximizations are provided in Section A.3. The final parameter estimate updates are

$$\begin{aligned} \alpha_{\hat{u}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W\left[x_{i\hat{u}}y_{iv}z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{n} \\ \beta_{\hat{v}\hat{u}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{k_C} E_W\left[x_{i\hat{u}}y_{i\hat{v}}z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W\left[x_{i\hat{u}}y_{iv}z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]} \\ \gamma_{\hat{j}\hat{v}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \sum_{u=1}^{k_A} E_W\left[x_{iu}y_{i\hat{v}}z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ j=1,\ldots,k_C}} E_W\left[x_{iu}y_{i\hat{v}}z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]} \\ \sigma_{\hat{j}\hat{d}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W\left[x_{iu}y_{iv}z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right](s_{i\hat{d}} - \eta_{10})}{(\eta_{11} - \eta_{10}) \sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W\left[x_{iu}y_{iv}z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]} \\ \mu_{\hat{j}\hat{d}\hat{r}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W\left[x_{iu}y_{iv}z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{\sum_{r=1}^{R} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} \sum_{i=1}^{n} E_Z\left[x_{iu}y_{iv}z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right][m_{i\hat{d}} = q_r]} . \end{aligned}$$

3.1.4 Semi-supervised EM

The EM algorithm can be adapted to account for partially-labeled data by noting that when the label is known for a particular isolate, the summation over the possible classification labels within the log-likelihood function simplifies to a single probability. As a result, that term of the full log-likelihood becomes tractable [23].

An equivalent way to consider the case of a partially-labeled data example is when you examine the complete log-likelihood (with the single-tier model presented here for simplicity):

$$\log L_C(\mathbf{\Theta}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \log \left\{ p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} \ .$$

In the case of labeled data, the normally-latent variable z_{ij} is actually observed, with $z_{ij} = 1$ when isolate i has class j. The rest of the EM computation proceeds normally, substituting the constant value 1 or 0 for z_{ij} as appropriate for labeled data.

For the multi-tier model, the same logic applies. The complete log-likelihood

$$\log L_C(\mathbf{\Theta}) = \sum_{i=1}^{n} \sum_{u=1}^{k_A} x_{iu} \sum_{v=1}^{k_B} y_{iv} \sum_{j=1}^{k_C} z_{ij} \log \left\{ p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\}$$

is simplified when one or more of x_{iu}, y_{iv}, and z_{ij} are actually known values instead of latent variables.

3.1.5 Modifications to EM Algorithm for Repeated Isolates

The update steps presented for each of the models in Section 3.1 assume that each data example in the data set is limited to a single identified strain of MTBC. Thus, if two strains are identified with the same spoligotype and MIRU measurements, they would have to be represented with different examples in the data set.

Alternatively, the EM updates can incorporate these repeated strains by including only one instance of a repeated isolate and using the repetition count information within the update step itself.

Let ℓ_i be the number of times isolate i is repeated in the database, and let $N = \sum_{i=1}^{n} \ell_i$ be the total number of all occurrences of all strains in the database. The log-likelihood of the models can be altered to include this informaion, causing changes to the parameter updates as well.

For example, for the multi-tier model, the log-likelihood becomes

$$\log L(\mathbf{\Theta}) = \sum_{i=1}^{n} \ell_i \log \left\{ \sum_{u=1}^{k_A} \alpha_u \sum_{v=1}^{k_B} \beta_{vu} \sum_{j=1}^{k_C} \gamma_{jv} p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) \right\}$$

which in turn causes the parameter updates to become

$$\begin{aligned}
\alpha_{\hat{u}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \ell_i \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{i\hat{u}} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right]}{N} \\
\beta_{\hat{v}\hat{u}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \ell_i \sum_{j=1}^{k_C} E_W \left[x_{i\hat{u}} y_{i\hat{v}} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right]}{\sum_{i=1}^{n} \ell_i \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{i\hat{u}} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right]} \\
\gamma_{\hat{j}\hat{v}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \ell_i \sum_{u=1}^{k_A} E_W \left[x_{iu} y_{i\hat{v}} z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right]}{\sum_{i=1}^{n} \ell_i \sum_{\substack{u=1,\ldots,k_A \\ j=1,\ldots,k_C}} E_W \left[x_{iu} y_{i\hat{v}} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right]} \\
\sigma_{\hat{j}\hat{d}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \ell_i \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W \left[x_{iu} y_{iv} z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right] (s_{i\hat{d}} - \eta_{10})}{(\eta_{11} - \eta_{10}) \sum_{i=1}^{n} \ell_i \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W \left[x_{iu} y_{iv} z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right]} \\
\mu_{\hat{j}\hat{d}\hat{r}}^{(t+1)} &= \frac{\sum_{i=1}^{n} \ell_i \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W \left[x_{iu} y_{iv} z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right]}{\sum_{r=1}^{R} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} \sum_{i=1}^{n} \ell_i E_Z \left[x_{iu} y_{iv} z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right] [m_{i\hat{d}} = q_r]} \; .
\end{aligned}$$

3.1.6 Modifications to EM Algorithm for Missing Data

As mentioned in Section 1.3.2, many of the isolates in the data set are missing one or more MIRU values. There are two possible ways of dealing with this missing data: by treating the missing MIRU values as additional unobserved latent variables and computing the expectation over these values as part of the EM algorithm, or—due to the condition independence between MIRU values given the class—the missing values can be ignored and the log-likelihood computed over the remaining values. In this thesis, as in [12], the latter approach is used, saving expensive latent variable computations at every iteration of the algorithm.

Thus, the computation of the MIRU probability given the class c_j

$$p(\mathbf{m}_i \mid c_j) = \prod_{d=1}^{D_M} p(m_{id} \mid c_j)$$

becomes instead the product of the probabilities for each d where m_{id} is not a missing MIRU value:

$$p(\mathbf{m}_i \mid c_j) = \prod_{d \in \left\{ \tilde{d} \,:\, m_{i\tilde{d}} \text{ is not missing} \right\}} p(m_{id} \mid c_j) \; .$$

3.1.7 Computing Maximum a Posteriori (MAP) Estimates

As described in [21], we can provide some regularization over the parameters being estimated in the EM algorithm by incorporating a prior distribution over these model parameters. This will have two benefits: it avoids overfitting to the data, and it can incorporate expert knowledge regarding these parameters.

When computing MAP estimates, instead of maximizing the likelihood of the data given the parameters, we are instead maximizing the mode of the posterior distribution of the parameters

$$p(\mathbf{\Theta} \mid S, M) \propto p(S, M \mid \mathbf{\Theta})p(\mathbf{\Theta})$$

The general maximization problem for EM will consider the function

$$\begin{aligned}\log\{p(\mathbf{\Theta} \mid S, M)\} &\propto \log p(S, M \mid \mathbf{\Theta})p(\mathbf{\Theta}) \\ &= \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} p(c_j)p(\mathbf{s}_i \mid c_j)p(\mathbf{m}_i \mid c_j) \right\} + \log\{p(\mathbf{\Theta})\}\end{aligned}$$

instead of the log-likelihood.

Adding a prior distribution $p(\mathbf{\Theta})$ to the models described in Chapter 2 is not a complex task. For the Bernoulli distributions used within the models to represent spoligotype spacer probabilities, the conjugate prior is the beta distribution. The two parameters of the beta distribution are often called *psuedo-counts* when used as a prior to the Bernoulli, as they can be thought of an initial number of "successes" and "failures" of the prior distritbution. In the case of the spoligotype spacer distributions, these psuedo-counts represent the prior knowledge of the number of isolates in a class that have the particular spacer present ("successes") or absent ("failures").

The conjugate prior to the categorical distribution (used for the class probabilties as well as the MIRU probabilities) is the Dirichlet distribution. Similarly to the beta distribution for the beta-Bernoulli model, the Dirichlet distribution in the Dirichlet-categorical model is parameterized by a set of pseudo-counts which can be thought of as the number of initial counts for each of the categories for with the Dirichlet distribution—and thus the categorical distribution—is defined.

Without demonstrating the entire revised calculation of the EM update steps (discussion of these calculations can be found in [21]), the end result of incorporating the Dirichlet and beta priors to these calculations involves a fairly simple midificaiton to the the update steps.

For example, let us consider the $\boldsymbol{\alpha} = \{\alpha_u\}_{u=1,\ldots,k_A}$ parameters with a Dirichlet prior $\boldsymbol{\alpha} \sim Dir(k_A, \tilde{\boldsymbol{\alpha}})$, where $\tilde{\boldsymbol{\alpha}} = \{\tilde{\alpha}_u\}_{u=1,\ldots,k_A}$ are the pseudo-counts for the prior of $\boldsymbol{\alpha}$. The EM-update step for the α_u parameter becomes:

$$\alpha_{\hat{u}}^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W\left[x_{i\hat{u}} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] + \tilde{\alpha}_{\hat{u}}}{n + \sum_{u=1}^{k_A} \tilde{\alpha}_u} \, .$$

Using similar notation for the β parameters, consider, for each $\hat{u} = 1, \ldots, k_A$, the parameters $\boldsymbol{\beta}_{\hat{u}} = \{\beta_{v\hat{u}}\}_{v=1,\ldots,k_B}$ with a Dirichlet prior $\boldsymbol{\beta}_{\hat{u}} \sim Dir(k_B, \tilde{\boldsymbol{\beta}}_{\hat{u}})$ where $\tilde{\boldsymbol{\beta}}u = \{\tilde{\beta}_{v\hat{u}}\}_{v=1,\ldots,k_B}$ are the pseudo-counts for the prior of $\boldsymbol{\beta}_{\hat{u}}$. The EM-update step for β_{vu} considering this prior is

$$\beta_{\hat{v}\hat{u}}^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k_C} E_W\left[x_{i\hat{u}} y_{i\hat{v}} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] + \tilde{\beta}_{\hat{v}\hat{u}}}{\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W\left[x_{i\hat{u}} y_{iv} z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] + \sum_{v=1}^{k_B} \tilde{\beta}_{v\hat{u}}} \, .$$

Following the same pattern, the γ_{jv} update becomes

$$\gamma_{\hat{j}\hat{v}}^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{u=1}^{k_A} E_W\left[x_{iu} y_{i\hat{v}} z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] + \tilde{\gamma}_{\hat{j}\hat{v}}}{\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ j=1,\ldots,k_C}} E_W\left[x_{iu} y_{i\hat{v}} z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] + \sum_{j=1}^{k_C} \tilde{\gamma}_{j\hat{v}}} \, .$$

For the σ parameters update, let us consider, for each $\hat{j} = 1, \ldots, k_C$ and $\hat{d} = 1, \ldots, D_S$, the parameter $\sigma_{\hat{j}\hat{d}}$ has a beta prior $\sigma_{\hat{j}\hat{d}} \sim Beta(\sigma_{jd+}, \sigma_{jd-})$ with the pseudo-counts σ_{jd+} and σ_{jd+}, where σ_{jd+} is the pseudo-count of class c_j having a spacer present in position d, and σ_{jd-} is the pseudo-count of class c_j not having a spacer present in position d. The EM-update step for σ_{jd} becomes

$$\sigma_{\hat{j}\hat{d}}^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W\left[x_{iu} y_{iv} z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] (s_{i\hat{d}} - \eta_{10}) + \sigma_{jd+}}{(\eta_{11} - \eta_{10}) \sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W\left[x_{iu} y_{iv} z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] + \sigma_{jd+} + \sigma_{jd-}} \, .$$

Finally, for the μ parameters, for each $\hat{j} = 1, \ldots, k_C$ and $\hat{d} = 1, .., D_M$, the parameters $\boldsymbol{\mu}_{\hat{j}\hat{d}} = \{\mu_{\hat{j}\hat{d}}r\}_{r=1,\ldots,R}$ have a Dirichlet prior $\boldsymbol{\mu}_{\hat{j}\hat{d}} \sim Dir(R, \tilde{\boldsymbol{\mu}}_{\hat{j}\hat{d}})$ where $\boldsymbol{\mu}\tilde{}_{\hat{j}\hat{d}} = \{\tilde{\mu}_{\hat{j}\hat{d}r}\}_{r=1,\ldots,R}$ are the pseudo-counts for the prior of $\boldsymbol{\mu}_{\hat{j}\hat{d}}$. The EM-update step for μ_{jdr} with this prior is

$$\mu_{\hat{j}\hat{d}\hat{r}}^{(t+1)} = \frac{\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W \left[x_{iu} y_{iv} z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right] + \tilde{\mu}_{\hat{j}\hat{d}\hat{r}}}{\sum_{r=1}^{R} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} \sum_{i=1}^{n} E_Z \left[x_{iu} y_{iv} z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right] [m_{i\hat{d}} = q_r] + \sum_{r}^{R} \tilde{\mu}_{\hat{j}\hat{d}r}} .$$

The choice of prior parameters (the pseudo-counts) can have a significant effect on the resulting model, depending on the amount of data being used to train the model. For the models used in this thesis, weak priors for the α, β, and γ distributions were set from historical data and models.

The priors for σ_{jd}, for each d, were set to the average of the counts in the database over all j. This ensures that if a spoligotype at a particular locus tends to have a certain value (either positive or negative) over all classes, the prior estimate will tend to that same value. Similarly, for μ_{jdr}, for each d the priors were set to the average of the counts in the database over all j. In the data set used, some MIRU loci values are very rare across all lineages, so using this type of prior ensures that even for classes with very few data points, those MIRU loci values tend to the MIRU loci values of other classes.

3.2 Model Initialization and Validation

The expectation-maximization algorithm requires a set of initial parameters which are then iteratively improved until convergence is met. However, the EM algorithm is only guaranteed to converge to a local maximum, and thus this choice of initial parameters has a significant impact on the final model parameters.

Section 3.2.1 describes the heuristic approach used to escape the local maximum. Section 3.2.2 describes a Monte Carlo cross-validation (MCCV) technique used to provide out-of-sample model validation metrics.

3.2.1 Model Initialization

The approach used in this project for dealing with the local maximum problem is random restarts. To find the best model, 100 randomly-initialized models are generated using the same training data set. Each of these is models is trained to convergence: a change in data point log-likelihood (the total data log likelihood divided by the number of data points in the data set) less than 10^{-6}, or a change in the norm of parameter weights less than 5×10^{-8}. The choice of 100 random restarts was chosen heuristically, as a number that provides an excellent chance of generating the best possible model; using any additional random restarts does not tend to increase the quality of the resulting model.

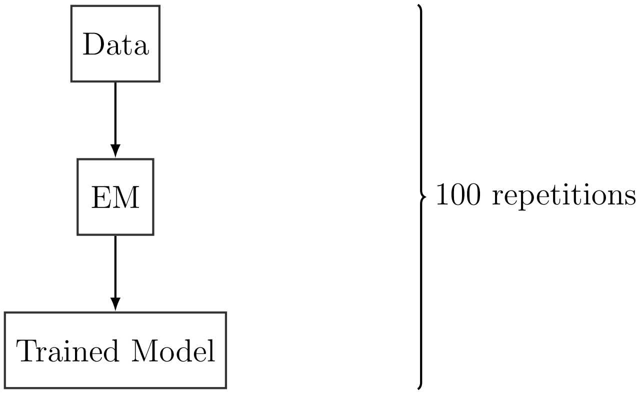


Figure 3.1: Schema for Random Repeat Initialization Approach used to Compute Final Model.

After training each model, the total data log-likelihood for each model is compared to each other and the model with the highest log-likelihood is selected.

3.2.2 Model Validation

For the purposes of validating our model generation approach, we employ a Monte Carlo cross-validation (MCCV) approach [24] to measure the data log-likelihood and classification accuracy of a test data set. In this approach, we repeatedly and randomly partition the data (sampling without replacement) into a training and testing data set, where the training data set contains 80% of the full data set and the test data set contains the remaining 20%.

For each of the data splits, we train the model using the training data set.

Similar to as described in 3.2.1, we must repeat the training process multiple times to ensure that the model being used for validation is not stuck in a local maximum.

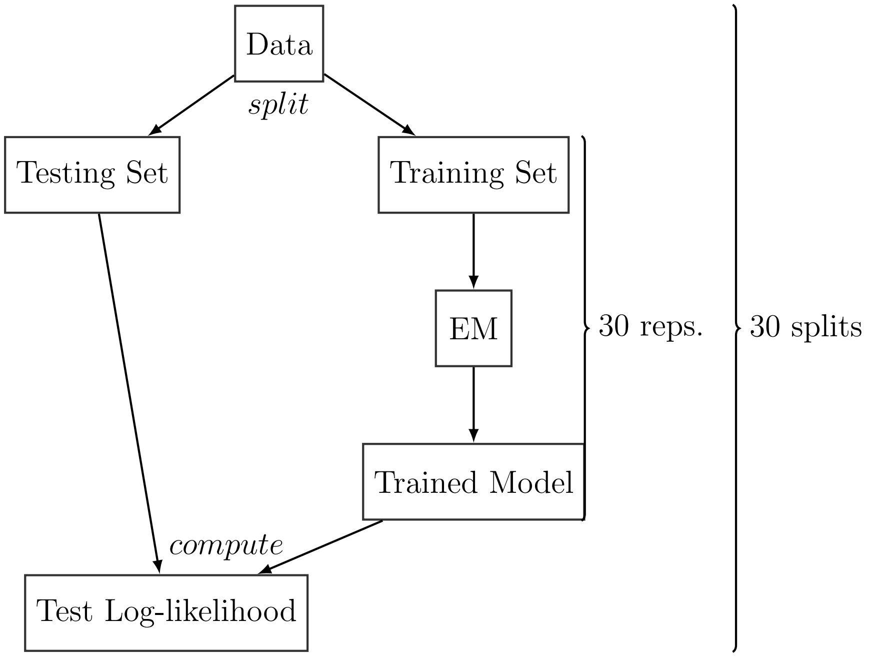


Figure 3.2: Schema for MCCV Approach used to Compute Out-of-sample Metrics.

Since the MCCV approach is used for testing the validity of our model generation approach and not for generating the best model, the full number of iterations used in generating a final model (100 repeats) is not required. When using the MCCV technique for comparing two models (for example, single-tier and multi-tier models), we used 30 repetitions of the randomly-initialized EM algorithm, and 30 separate data splits.

Once the models are trained, the model is then used to classify the test data set. Once classified, we can generate a confusion matrix between initial labels (if labeled) and labels as classified by the trained model, which we can use to compute accuracy measurements. We can also compute the test data log-likelihood. These metrics were then either compared, for each split, against competing models, or averaged over all the splits for the final out-of-sample metric.

As an alternative method to comparing models using out-of-sample test metrics, Section 3.3 describes a model selection process used to systematically compare different models and find the best model among several models with differing numbers of model parameters.

3.3 Model Selection

In the semi-supervised scenario, our algorithm is initialized with the same number of classification labels—either k labels in the single-tier model, or k_A, k_B, and k_C labels in the multi-tier model—as exists in the data set. However, given that many of the unlabeled data points may not appropriately fit in these existing labels, or even in the same that some of the labeled data points would benefit by being split off into a new classification, some extra flexibility to this number of classification labels is required.

Unfortunately, it is not clear how many extra classification labels are required to fit the data. One possible method to discover this "correct" number of labels would be to use the MCCV approach described in Section 3.2.2, and compare test log-likelihood values between models with varying numbers of extra classification labels. However, the MCCV approach requires several splits, and several repetitions per split, which takes an excessive amount of time.

To avoid the excessive amount of time required for the MCCV approach, we instead consider an approach based on the Akaike information criterion (AIC) [25] or the Bayesian information criterion (BIC) [26]. On the MTBC data set, both of these criteria result in identical conclusions, and thus we concentrate on the AIC.

The formula for the AIC is

$$AIC = 2k - 2\log(L)$$

where L is the maximum log-likelihood of the data given the model and k is the number of free parameter in the model. At its most basic level, the AIC is simply the log-likelihood penalized by a measurement of the complexity of the model, thus preventing model overfitting. Between two models, the model with the *smaller* AIC should be chosen.

The AIC is used to compare different models using the following approach: for each of the possible numbers of extra classification labels, we train a model using the entire data set. As in Section 3.2.1 and 3.2.2, to avoid getting stuck in a local maximum, we repeat the training of this model (for each of the possible numbers of extra classification numbers) multiple times and only calculate the AIC for the model with the best log-likelihood.

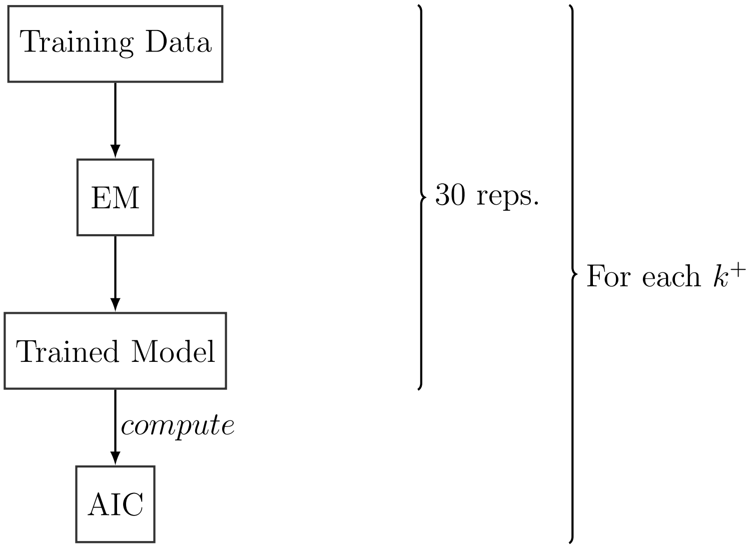


Figure 3.3: Schema for Model Selection Approach. The variable k^+ represents the various possible numbers of extra classification labels.

In the case of the multi-tier models, we must examine the possibility of adding extra top-level, mid-level, and sub-level labels. To find the best number of extra labels at these various levels, a grid search is performed over all the possible combinations, and the combination that provides the best AIC value is selected as the best model to use.

Since this type of grid search over a three-dimensional input space can computationally expensive, we limit the number of random-repeat repetitions used for training a model: in our experiments, repeating 20 times provides a good trade-off between speed of execution and the local maximum avoidance.

For the single-tier models, let k^+ be the number of extra labels. The model selection search was performed over the range $k^+ = [0, 24]$ (with every value between 0 and 24 attempted). For the multi-tier models, let k_A^+ be the number of extra top-level labels, k_B^+ be the number of extra mid-level labels, and k_C^+ be the number of extra sub-level labels. The grid search was performed over the ranges $k_A^+ = [0, 3]$, $k_B^+ = [0, 8]$, and $k_C^+ = [0, 24]$. These ranges were chosen as a result of initial experimentation, and noticing that trying numbers of extra labels beyond these ranges consistently provided lower AIC values.

CHAPTER 4
Results

In this section we examine the effectiveness of the multi-tier model. We begin by comparing the single-tier hidden-parent model (described in Section 2.4) against the multi-tier hidden-parent model (described in Section 2.5). Given the benefits of the hidden parent approach demonstrated by SPOTCLUST [11], we did not include a model without hidden parent nodes in our comparison. The description of the single-tier model without hidden parents provided in Section 2.3 is provided instead as an simple demonstration of the general modeling approach that is used in the hidden parent models.

Our secondary experiment compares the spoligotype-only approach similar to that used in [11] and the proposed spoligotype / MIRU approach discussed in Chapter 2.

The data set used for compiling these results is presented and discussed in Section 1.3.

Section 4.1 discusses the results of the comparisons of these models. Section 4.2 describes the details of the best probability model resulting from these comparisons and provides some discussion of these results.

4.1 Experimental Results

We investigate four types of models: single-tier spoligotype-only, single-tier with MIRU, multi-tier spoligotype-only, and multi-tier with MIRU. As described in Section 3.3, each went through a model selection process wherein the best number of extra labels was found for each. These results of these model selection efforts are found in Section 4.1.1. After the model selection process, the selected model for each model type being analyzed underwent Monte Carlo cross-validation using the same training / testing data splits for each model type. The resulting test log-likelihood and accuracy measurements are compared in Section 4.1.2.

4.1.1 Model Selection Results

Since the data set consists of a significant number of unlabeled or partially-labeled data examples, there exists a possibility that the existing number of labels (presented in Tables1.1 and 1.2) are not sufficient for the data set. Some isolates may not "fit" well into existing classification labels, and may require new labels to be created in order to maximize the effectiveness of the model. Therefore, models with extra labels were considered.

The model selection process used to determine how many extra labels to use is discussed in Section 3.3. This method—which uses the Akaike Information Criterion (AIC) to regulate the trade-off between goodness-of-fit of a model based on the training data and the model's complexity—provides the results found in this section.

Table 4.1 shows the results for finding an appropriate number of extra labels for single-tier models. In this table, k^+ represents the best additional number of labels used by the model, as determined by comparing the AIC of several different models of varying size.

Table 4.1: Model Selection Results for Single-tier Model Types.

Model Type	k^+
Single-tier spoligotype-only	3
Single-tier spoligotype & MIRU	6

Figures 4.1 and 4.2 show the plots of the model selection results for the single-tier spoligotype-only and spoligotype & MIRU models. These graphs show the negative AIC value plotted against the different possible numbers of extra labels, and is averaged over 15 repetitions of the model selection process, with the standard deviation shown on the plot with error bars. Note that a larger parameter space was examined than is shown in these graphs; these graphs only show the areas of the parameter space with the highest negative AIC value. The full parameter space examined is discussed in Section 3.3.

Table 4.2 shows the results for finding an appropriate number of extra labels for multi-tier models. In this table, k_A^+ represents the additional number of top-level

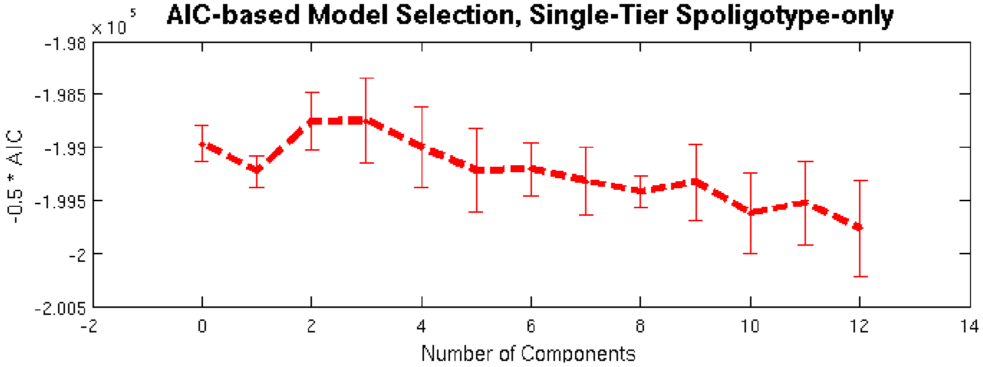


Figure 4.1: Model Selection Plot for Single-Tier Spoligotype-only Model.

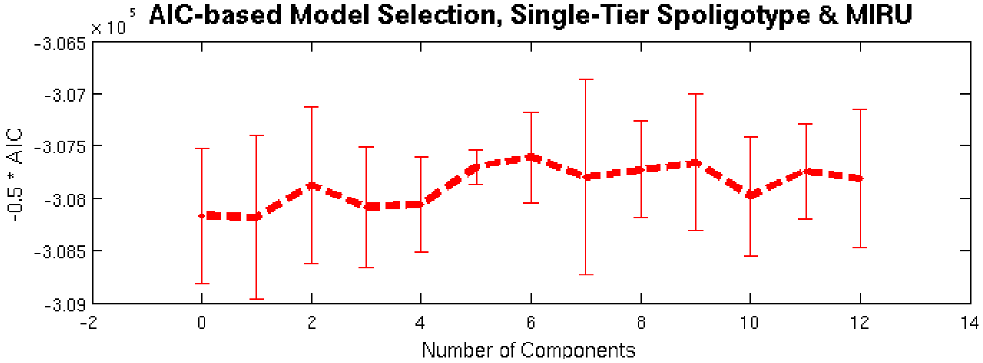


Figure 4.2: Model Selection Plot for Single-Tier Spoligotype & MIRU Model.

labels, k_B^+ represents the additional number of mid-level labels, and k_C^+ represents the additional number of sub-level labels used by the model. As with the single-tier model selection, the values of k_A^+, k_B^+, and k_C^+ were determined by comparing the AIC of several different models of varying size, as discussed in Section 3.3.

In both Table 4.1 and Table 4.2, the spoligotype & MIRU model uses more extra labels than is spoligotype-only counterpart. This fact indicates that the inclusion of MIRU data splits the data set into more lineage classifications, according to the model, than considering spoligotype alone. Since the existing lineage hierarchy is a result of predominantly spoligotype-only analyses, the requirement of additional classification labels as a result of MIRU analysis is unsurprising.

Figures 4.3 and 4.4 show the plots of the model selection results for the multi-tier spoligotype-only and spoligotype & MIRU models. These graphs show the negative AIC value plotted against the different possible numbers of extra labels,

Table 4.2: Model Selection Results for Multi-tier Model Types.

Model Type	k_A^+	k_B^+	k_C^+
Multi-tier spoligotype-only	0	0	2
Multi-tier spoligotype & MIRU	0	0	6

and is averaged over 15 repetitions of the model selection process, with standard deviation shown as error bars. Each set of plots shows the one-dimensional curve of the AIC value; in actuality, the model selection process occurred over the three-dimensional model selection space (the three dimensions being the extra number of top-level, mid-level, and sub-level labels). Additionally, note that a larger parameter space than is shown in these graphs was examined; these graphs only show the areas of the parameter space with the highest negative AIC values. The grid over which the entire parameter space was search is discussed in Section 3.3.

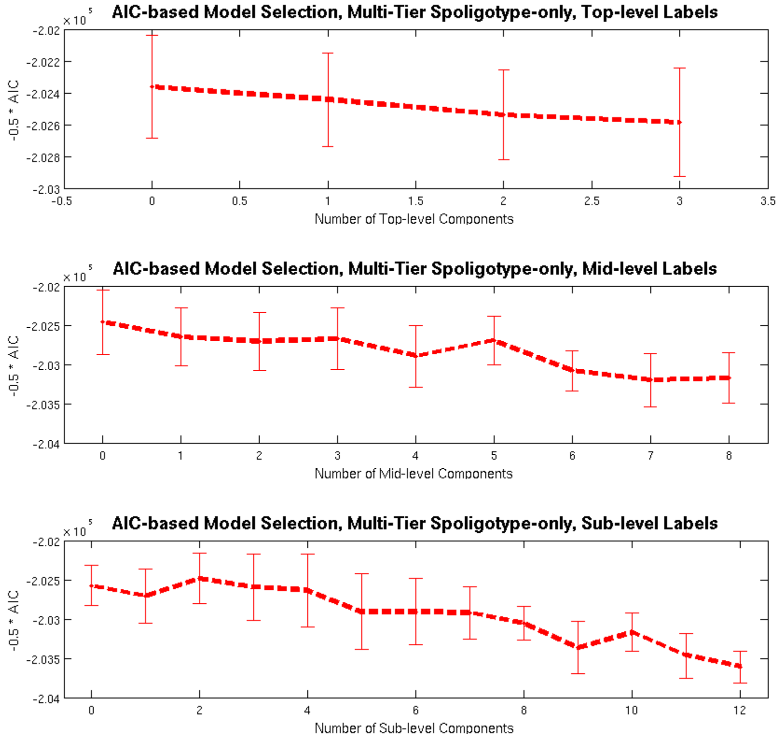


Figure 4.3: Model Selection Plots for Multi-Tier Spoligotype-only Model.

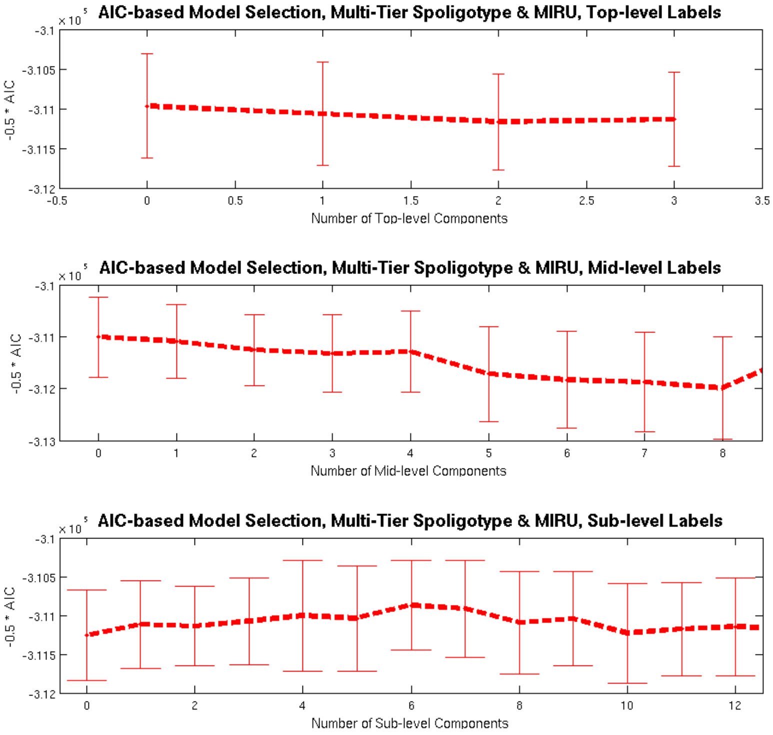


Figure 4.4: Model Selection Plots for Multi-Tier Spoligotype & MIRU Model.

4.1.2 Model Comparison Results

Once an appropriate model is selected for each of the four model styles, these model styles can then be compared to determine the best model to use for effective modeling of this data. These comparisons all use the cross-validation techniques described in Section 3.2.2

For the purposes of comparison, we must consider what metrics are best suited when comparing the different models. The single-tier and multi-tier models cannot be adequately compared via simple accuracy measurements, as accuracy measurements typically only check if the actual label and predicted labels are the same. Since the multi-tier models have three different labels, there can varying degrees of “correctness” and “incorrectness” for a particular predicted label. Thus, to compare single-tier and multi-tier models, we consider the cross-validated test log-likelihood between the two models. This comparison is presented in Section 4.1.2.1.

When comparing the spoligotype-only and spoligotype & MIRU versions of a model—either single-tier or multi-tier—the test log-likelihood does not provide a valid comparison metric. Since the spoligotype-only models do not take into account any of the MIRU data, any computed test log-likelihood using MIRU data for the combined spoligotype & MIRU model would be significantly smaller (as a result of multiplying by the probability that each of the extra MIRU values occur). However, comparing the classification accuracy of these two models does provide a valid comparison metric: the single-tier spoligotype-only model can be compared to the single-tier spoligotype & MIRU model, and the multi-tier spoligotype-only model can be adequately compared to the spoligotype & MIRU model. This comparison is presented in Section 4.1.2.2.

4.1.2.1 Multi-tier vs. Single-tier Models

Table 4.3 shows the test log-likelihood results comparing the single-tier and multi-tier spoligotype & MIRU models, based on 60 splits of the data.

Table 4.3: Cross-validated Test Log-likelihood (Average ± Standard Deviation) for Spoligotype & MIRU models.

Model	Test Log-likelihood
Multi-tier Spoligotype & MIRU	$\mathbf{-28752 \pm 289}$
Single-tier Spoligotype & MIRU	-28901 ± 289

Table 4.3 shows the test log-likelihood results comparing the single-tier and multi-tier spoligotype-only, based on 60 splits of the data.

Table 4.4: Cross-validated Test Log-likelihood (Average ± Standard Deviation) for Spoligotype-only Models.

Model	Test Log-likelihood
Multi-tier Spoligotype-only	$\mathbf{-17581 \pm 207}$
Single-tier Spoligotype-only	-17659 ± 180

To determine if the difference in the test log-likelihood is statistically significant, a paired t-test was performed between multi-tier and single-tier results for both spoligotype & MIRU and spoligotype-only models, with the results presented in Table 4.5. For each comparison, the average difference between two log-likelihoods is presented, along with the p-value of a one-tailed paired t-test, which indicates in both cases that the multi-tier model provides a significantly better cross-validated log-likelihood than the single-tier.

Table 4.5: Model Comparison Results with Average Improvement in Cross-validated Test Log-likelihood and p-values from Paired t-Test, Multi-tier vs. Single-tier.

Model A	Model B	LL Diff.	p-value
Multi-tier Sp.&MIRU	Single-tier Sp.&MIRU	148.94 ± 121.11	7.69×10^{-14}
Multi-tier Sp.-only	Single-tier Sp.-only	77.49 ± 93.22	1.20×10^{-8}

4.1.2.2 Spoligotype & MIRU vs. Spoligotype-only models

Table 4.6 shows the cross-validated classification accuracy results comparing the multi-tier spoligotype & MIRU and spoligotype-only models, based on 60 splits of the data.

Table 4.6: Cross-validated Classification Accuracy (Average ± Standard Deviation) for Multi-tier Models.

Model	Classification Accuracy
Multi-tier Spoligotype & MIRU	$\mathbf{91.97 \pm 0.031}$
Multi-tier Spoligotype-only	91.53 ± 0.028

Table 4.7 shows the cross-validated classification accuracy results comparing the single-tier spoligotype & MIRU and spoligotype-only models, based on 60 splits of the data.

To determine if the difference in the cross-validated classification accuracy is statistically significant, a paired t-test was performed between spoligotype & MIRU and spoligotype-only results for both single-tier and multi-tier models, with the results presented in Table 4.8. For each comparison, the average difference between two classification accuracies is presented, along with the p-value of a one-tailed paired t-test. These p-values indicate that, in both cases, the spoligotype & MIRU

Table 4.7: Cross-validated Classification Accuracy (Average ± Standard Deviation) for Single-tier Models.

Model	Classification Accuracy
Single-tier Spoligotype & MIRU	$\mathbf{91.71 \pm 0.031}$
Single-tier Spoligotype-only	90.79 ± 0.040

model performs better than the spoligotype-only model in term of classification accuracy, at a 0.05 significance level.

Table 4.8: Model Comparison Results with Average Improvement in Cross-validated Classification Accuracy and p-values from Paired t-Test, MIRU & Spoligotype vs. Spoligotype-only.

Model A	Model B	Acc. Diff.	p-value
Multi-tier Sp.&MIRU	Multi-tier Sp.-only	0.00432 ± 0.0143	0.0112
Single-tier Sp.&MIRU	Single-tier Sp.-only	0.00916 ± 0.0322	0.0158

Based on the results in Tables 4.5 and 4.8, the multi-tier model that takes advantage of both the MIRU and spoligotype data both provides a higher cross-validated test log-likelihood than the single-tier model, and provides a more accurate classifier than the spoligotype-only model, and thus is the best candidate model among these tested models.

4.2 Resulting Probabilistic Lineage Model

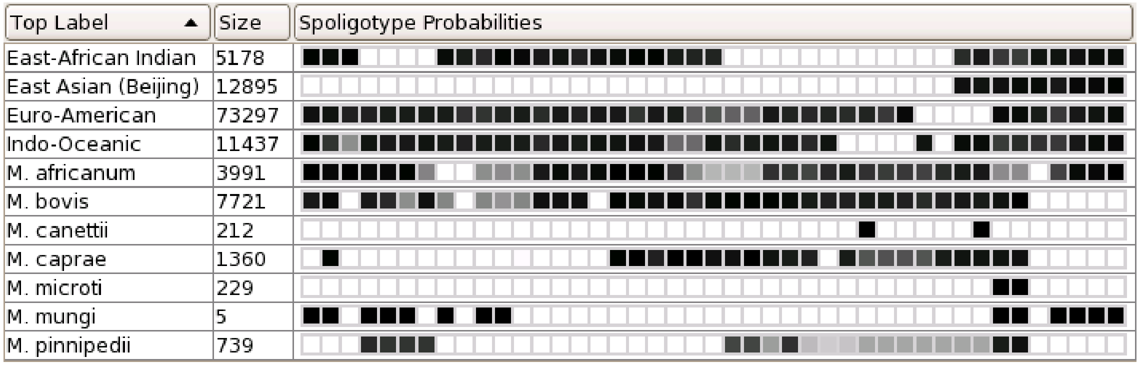
After reviewing the model comparison statistics, the three-tier spoligotype & MIRU model type using six extra sub-level labels proves to be the model type which most effectively effectively models the data. Using the full model training process described in Section 3.2.1 on the full combined data set, a final model was created.

Figure 4.9 shows the original data labels (Figure 4.9a) and predicted data labels (Figure 4.9b) aggregated to only show the results for the combined top-level lineage classification. Each figure shows the average (for the original data) or prototype (for the predicted labels) spoligotypes and MIRU for these top-level classes.

Figure 4.10 breaks down the labels in Figure 4.9 to show the the original data labels (Figure 4.10a) and predicted data labels (Figure 4.10b) aggregated to show the mid-level lineage classification.

Figures 4.11 and 4.12 break down these results even further, displaying the original sub-level label information, which can be compared to the predicted sub-level labels in Figures 4.13 and 4.14.

The spoligotype column in these figures shows the chance that a spoligotype spacer occurs in each of the 43 possible locations, with a black box indicated a near-certain probability of spacer, and a white box meaning a near-certain probability of no spacer existing. Figure 4.5 shows an example which displays the spoligotype probabilities from the data set, aggregated over the top-level labels. From this figure one can see that, for example, the first three spoilgotype spacers for East-African Indian lineage have a very high presence probability, while the next four spacers are more likely to be absent. As an another example, for the Indo-Oceanic lineage, the third spoligotype spacer has a fairly even probability of being either absent or present. Figure 4.6 provides a legend for the various shades of gray shown for the spoligotype spacers.



Top Label ▲	Size	Spoligotype Probabilities
East-African Indian	5178	
East Asian (Beijing)	12895	
Euro-American	73297	
Indo-Oceanic	11437	
M. africanum	3991	
M. bovis	7721	
M. canettii	212	
M. caprae	1360	
M. microti	229	
M. mungi	5	
M. pinnipedii	739	

Figure 4.5: Spoligotype Label Example.

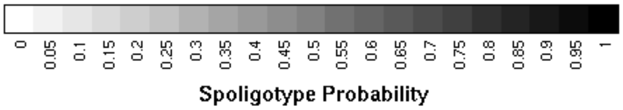
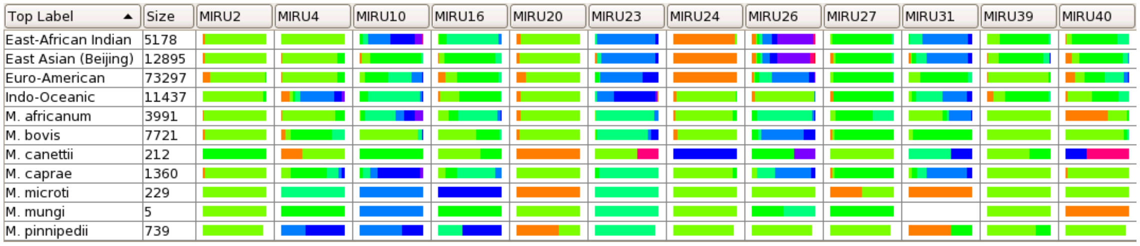


Figure 4.6: Spoligotype Spacer Data Legend.

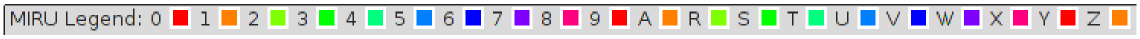
The MIRU columns in these figures give a visual impression of the categorical probabilities for each of the 12 MIRU locations for an MTBC isolate. Figure 4.7 shows an example of the MIRU probabilities from the data set, aggregated over the top-level labels. In each column a color bar is displayed which represents the full probability distribution for that MIRU locus. The portion of the bar which contains a particular color indicates the probability that the MIRU locus has the value corresponding to that color as provided by the legend shown in Figure 4.8.

As an example, if we consider MIRU24 values, the East African Indian, East Asian (Beijing), and Euro-American lineages have a very high probability of the MIRU24 value being equal to 1. The remaining lineages have a high probability of the MIRU24 value instead being equal to 2, or in the case of *M. canetti*, the MIRU24 value is equal to 6. This corresponds to the distinction between modern and ancestral lineages, and in fact the MIRU24 value is often looked to as an indicator of the ancestral-modern distinction. Note that while some colors due repeat on the MIRU legend in Figure 4.8, for the most part, the R-Z MIRU values are fairly rare and do not often constitute a significant probability on these visualizations.



Top Label ▲	Size	MIRU2	MIRU4	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40
East-African Indian	5178												
East Asian (Beijing)	12895												
Euro-American	73297												
Indo-Oceanic	11437												
M. africanum	3991												
M. bovis	7721												
M. canettii	212												
M. caprae	1360												
M. microti	229												
M. mungi	5												
M. pinnipedii	739												

Figure 4.7: MIRU Label Example.



MIRU Legend: 0 🟥 1 🟧 2 🟩 3 🟩 4 🟩 5 🟦 6 🟦 7 🟪 8 🟥 9 🟥 A 🟧 R 🟩 S 🟩 T 🟩 U 🟦 V 🟦 W 🟪 X 🟥 Y 🟥 Z 🟧

Figure 4.8: MIRU Categorical Data Legend.

The lineage results figures also contain the size information for each of the labels (original or predicted). This number simply refers to the number of isolates in each of the lineage classifications (aggregated over lower-level lineages, in the case of the top-level and mid-level figures).

For the predicted label figures (Figures 4.9b, 4.10b, 4.13, and 4.14), the "Prob." column represents the probability for the specified label (in the column proceeding it). For instance, in Figure 4.9b, the "Prob." column indicates the base probability that corresponding top-level label is chosen in the model—the value α_u from Section 2.5. The values in the this column in Figure 4.9b sum to 1.

For the predicted label figured with multiple label levels, such as Figure 4.10b, the "Prob." column following a mid-level label represents the probability of the mid-level label given the specified top-level label—the value β_{vu} from Section 2.5. For each unique top-level label, the values in the mid-level probability column should sum to 1, and the values for unique top-level labels in the top-level probability column again sum to 1. For example, in Figure 4.10b, if we consider the *M. africanum* rows, the chance for an isolate in *M. africanum* to be classified as "West African 1" is 0.578, and the chance for an isolate to be classified as "West African 2" is 0.422.

In a similar fashion, in the sub-level predicted label figures (Figure 4.13 and Figure 4.14), the "Prob." column following the sub-level label represents the probability of the sub-level label given the specified mid-level label—the value γ_{jv} from Section 2.5. These sub-level probability values sum to 1 for each mid-level label.

Figures 4.15 and 4.15 show the confusion matrices between the original data set and the same data set as predicted by the model. These tables are provided to display the isolates on which the model disagrees with the data set, and to show which labels are assigned to those isolates that are originally unlabeled.

In Figures 4.15 and 4.16, some of the incorrect labelings are worthy of further discussion.

Specifically, the ancestral (Indo-Oceanic, *M. africanum*, *M. bovis*, *M. canetti*, *M. caprae*, *M. microti*, *M. mungi*, and *M. pinnipedii*) and the modern (Beijing, East-African Indian, and Euro-American) lineages should be well-delineated. And while the isolates labeled as modern lineages in the original data set normally are not classified as ancestral, there are some isolates in the Indo-Oceanic and *M. africanum* families which are being classified as Euro-American of East-African Indian. By looking at the mid-level confusion matrix in Figure 4.16, we can see that a large portion of these misclassified isolates stem from two separate misclassifications: some isolates originally labeled as Indo-Oceanic/Bangladesh (top-level label as Indo-Oceanic and mid-level label as Bangladesh) were misclassified as Euro-American/X (69 isolates), and some isolates that were only labeled as Indo-Oceanic at the top-level and were missing a mid-level label were classified by the model as East-African Indian (46 isolates).

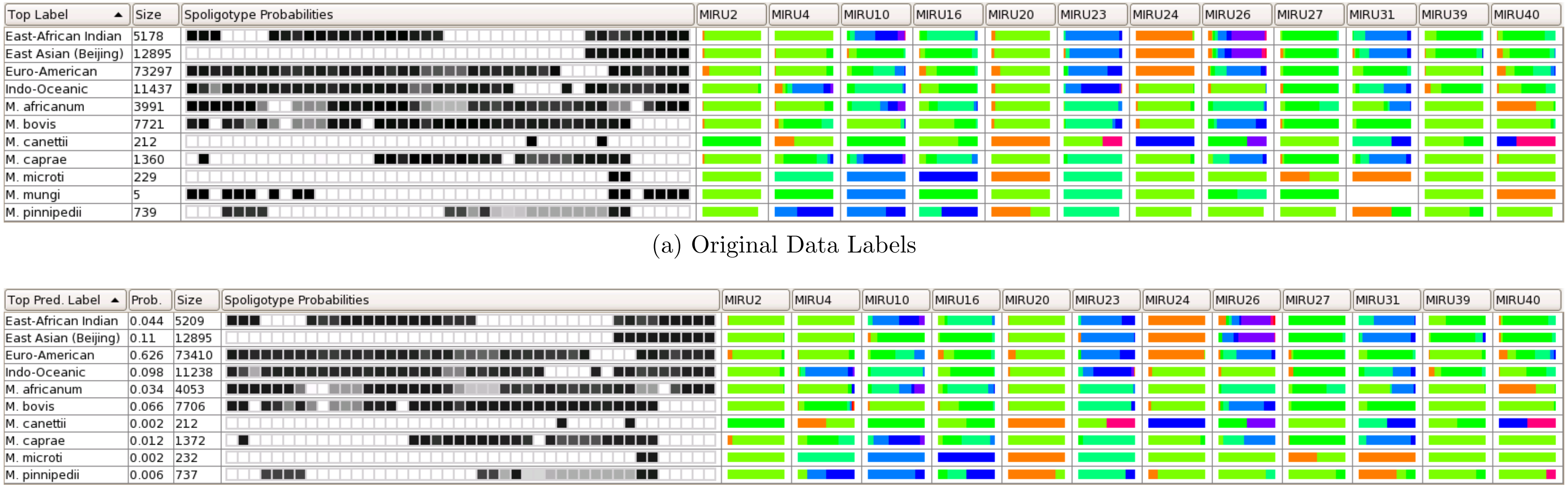
Upon further analysis of the misclassified Indo-Oceanic/Bangladesh isolates by reviewing the sub-level label confusion matrix (which are not included in this thesis for space) and by investigating the individual data examples, the reason for this particular mistake is due to the fact that the Indo-Oceanic/Bangladesh isolates that were misclassified did not possess MIRU data, and their spoligotype data was fairly close to that of some spoilgotype patterns in the Euro-American/X lineage. Examining Figure 4.13, the EAI7-BGD2 sub-level label has moved from Indo-Oceanic/Bangladesh/EAI7-BGD2 to Euro-American/X/EAI7-BGD2. The isolates labeled as EAI7-BGD2 had more spoligotype similarity to many isolates labeled in Euro-American and thus were moved there by the model This indicates an inconsistency in the labeling between IP (who provided isolates with EAI7-BGD2 labels) and CDC (who provided the isolates with Euro-American-labels), which the model resolves in favor of classifying these isolates as Euro-American/X.

Reviewing the misclassified Indo-Oceanic labels that were classified as East-African Indian, the mislabeled data either had missing MIRU data or had both MIRU and spoligotype data that closely resembled East-African Indian isolates. Given the fact that these labels were only classified at the top-level as Indo-Oceanic, it is possible that these isolates are in fact mislabeled in the data set.

Another interesting discrepancy between the original data labels and the predicted labels is the fact that the *M. mungi* top-level and mid-level label is not used at all in the predicted model. Given the very small number of the *M. mungi* isolates in the database, it appears that the model ignored the top-level and mid-level *M. mungi* labels entirely, and East-African Indian. Given more *M. mungi* data, the model training process may not ignore the *M. mungi* isolates, and this misclassification may not occur. Since the *M. mungi* data is so rare in the database, but the classification of *M. mungi* as a separate lineage is strongly supported, an alternative would be to artificially upweight the *M. mungi* isolates in the data set specifying a higher number of repetitions for each isolate.

This reveals another interesting consequence of the model design: since the structure of the three-tier model allows for any top-level label to be a parent of any mid-level label, and any mid-level label to be a parent of any sub-level label, it is possible for a label that, in the original data set, to switch to another branch of the hierarchy. For example, as can be seen in Figures 4.13 and 4.14, the *M. mungi* sub-level classification label ended up moving through the hierarchy under Indo-Oceanic/India. Similarly, the EAI7-BGD2 label ended up in the Euro-American/X hierarchy (a side effect of the confusion discussed above). Obviously, the names of these labels are no longer valid for their new location in the hierarchy; instead, they indicate areas in which the model had a higher log-likelihood when an extra sub-level classification exists at that location in the hierarchy. Thus, while the model was trained with only six extra sub-level classification labels, if we include *M. mungi* and EAI7-BGD2, eight new locations within the hierarchy were used.

One last point of discussion regarding the predictive model is the final disposition of the extra sub-level labels the model was provided. Examining Figures 4.13 and 4.14, the new labels (called 'OtherSub1',. . . ,'OtherSub6') extended the existing label hierarchy as follows: two news sub-level labels were added to the East-African Indian lineage, one sub-level label was added to the Euro-American/Haarlem lineage, one sub-level label was added to the Euro-American/LAM lineage, and two sub-level labels were added to the Euro-American/X lineage. Examining the spoligotype and MIRU prototypes for these new sub-level labels in Figures 4.13 and 4.14, it appears that these new sublineage labels have definite distinct spoligotype and MIRU patterns, and may provide new sub-level classifications which do not currently exist in the labeled data set. As discussed in Section 4.1.1, the inclusion of MIRU in the model selection process indicates that more sublineages may exist than previously identified using spoligotype-only models.



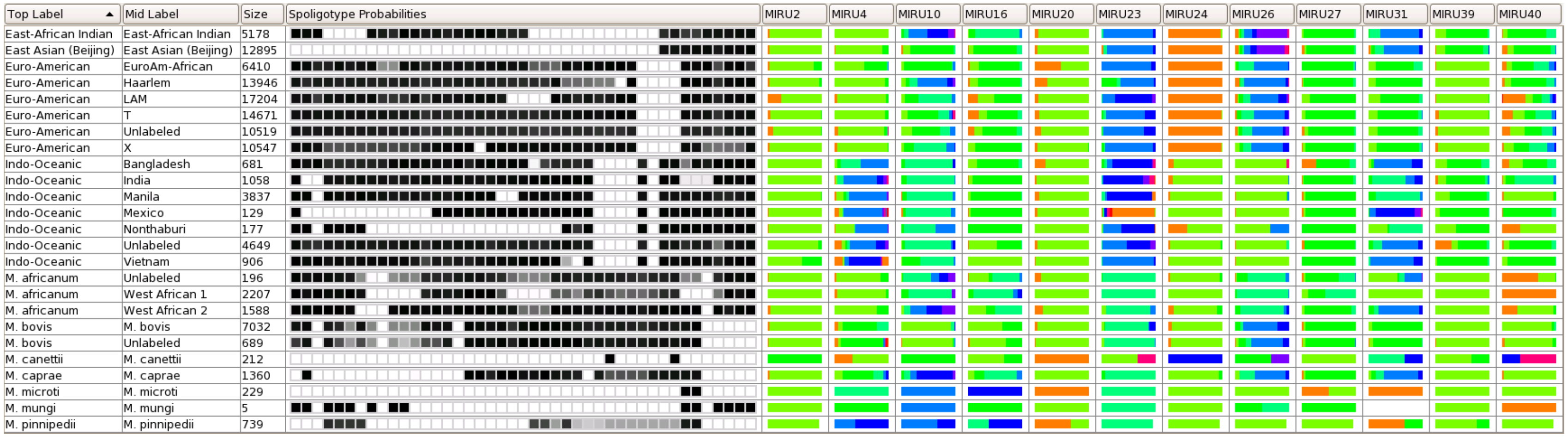
Top Label ▲	Size	Spoligotype Probabilities	MIRU2	MIRU4	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40
East-African Indian	5178													
East Asian (Beijing)	12895													
Euro-American	73297													
Indo-Oceanic	11437													
M. africanum	3991													
M. bovis	7721													
M. canettii	212													
M. caprae	1360													
M. microti	229													
M. mungi	5													
M. pinnipedii	739													

(a) Original Data Labels

Top Pred. Label ▲	Prob.	Size	Spoligotype Probabilities	MIRU2	MIRU4	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40
East-African Indian	0.044	5209													
East Asian (Beijing)	0.11	12895													
Euro-American	0.626	73410													
Indo-Oceanic	0.098	11238													
M. africanum	0.034	4053													
M. bovis	0.066	7706													
M. canettii	0.002	212													
M. caprae	0.012	1372													
M. microti	0.002	232													
M. pinnipedii	0.006	737													

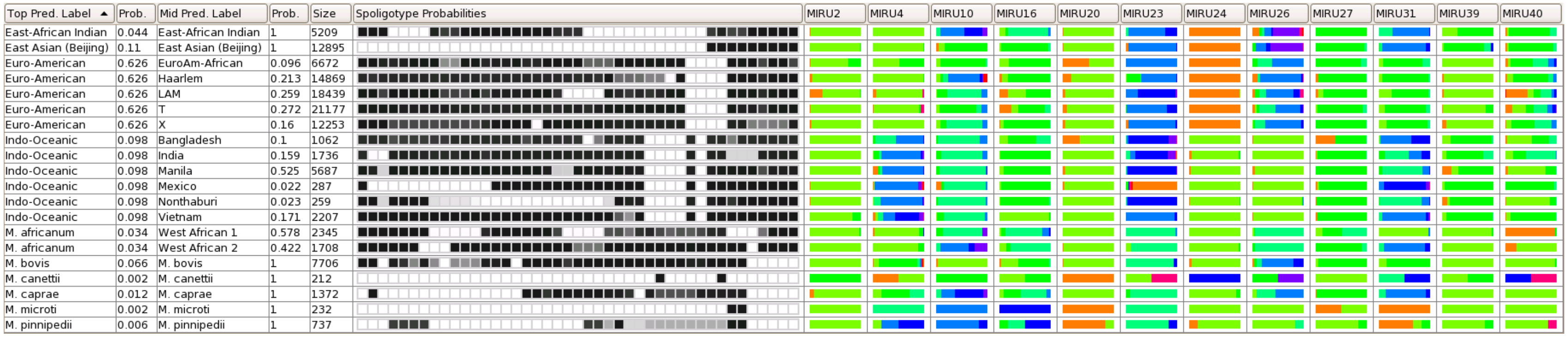
(b) Predicted Labels

Figure 4.9: Original and Predicted Top-level Data Labels, Final Three-Tier Model. Figure 4.9a shows the original data aggregated to only display the top-level labels, along with the top-level class sizes and the average spoligotype and MIRU values. Figure 4.9b shows the top-level labels as predicted by the model, with the probability for each class, the predicted class size (when applied to the data set), and the prototypes for spoligotypes and MIRU values.



Top Label ▲	Mid Label	Size	Spoligotype Probabilities	MIRU2	MIRU4	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40
East-African Indian	East-African Indian	5178													
East Asian (Beijing)	East Asian (Beijing)	12895													
Euro-American	EuroAm-African	6410													
Euro-American	Haarlem	13946													
Euro-American	LAM	17204													
Euro-American	T	14671													
Euro-American	Unlabeled	10519													
Euro-American	X	10547													
Indo-Oceanic	Bangladesh	681													
Indo-Oceanic	India	1058													
Indo-Oceanic	Manila	3837													
Indo-Oceanic	Mexico	129													
Indo-Oceanic	Nonthaburi	177													
Indo-Oceanic	Unlabeled	4649													
Indo-Oceanic	Vietnam	906													
M. africanum	Unlabeled	196													
M. africanum	West African 1	2207													
M. africanum	West African 2	1588													
M. bovis	M. bovis	7032													
M. bovis	Unlabeled	689													
M. canettii	M. canettii	212													
M. caprae	M. caprae	1360													
M. microti	M. microti	229													
M. mungi	M. mungi	5													
M. pinnipedii	M. pinnipedii	739													

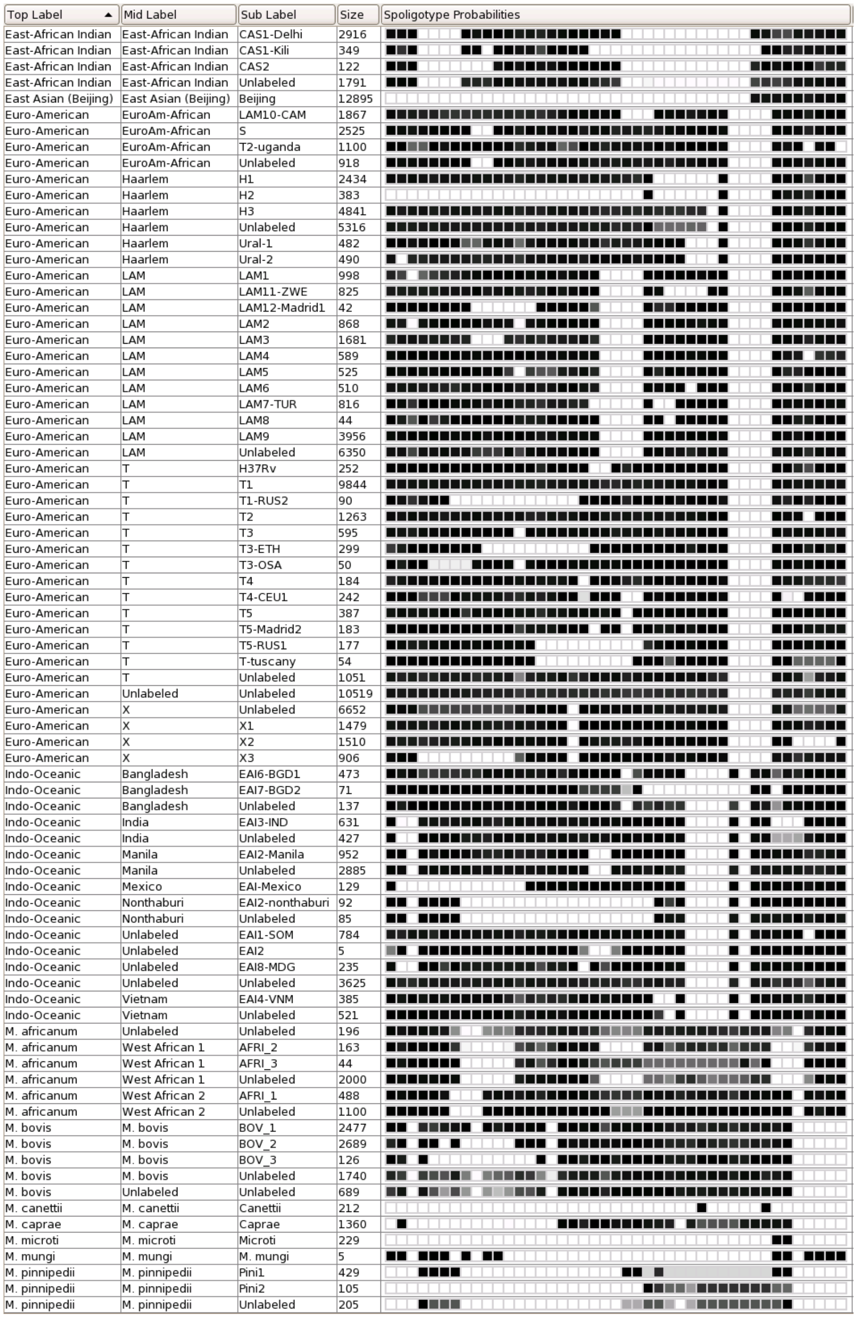
(a) Original Data Labels



Top Pred. Label ▲	Prob.	Mid Pred. Label	Prob.	Size	Spoligotype Probabilities	MIRU2	MIRU4	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40
East-African Indian	0.044	East-African Indian	1	5209													
East Asian (Beijing)	0.11	East Asian (Beijing)	1	12895													
Euro-American	0.626	EuroAm-African	0.096	6672													
Euro-American	0.626	Haarlem	0.213	14869													
Euro-American	0.626	LAM	0.259	18439													
Euro-American	0.626	T	0.272	21177													
Euro-American	0.626	X	0.16	12253													
Indo-Oceanic	0.098	Bangladesh	0.1	1062													
Indo-Oceanic	0.098	India	0.159	1736													
Indo-Oceanic	0.098	Manila	0.525	5687													
Indo-Oceanic	0.098	Mexico	0.022	287													
Indo-Oceanic	0.098	Nonthaburi	0.023	259													
Indo-Oceanic	0.098	Vietnam	0.171	2207													
M. africanum	0.034	West African 1	0.578	2345													
M. africanum	0.034	West African 2	0.422	1708													
M. bovis	0.066	M. bovis	1	7706													
M. canettii	0.002	M. canettii	1	212													
M. caprae	0.012	M. caprae	1	1372													
M. microti	0.002	M. microti	1	232													
M. pinnipedii	0.006	M. pinnipedii	1	737													

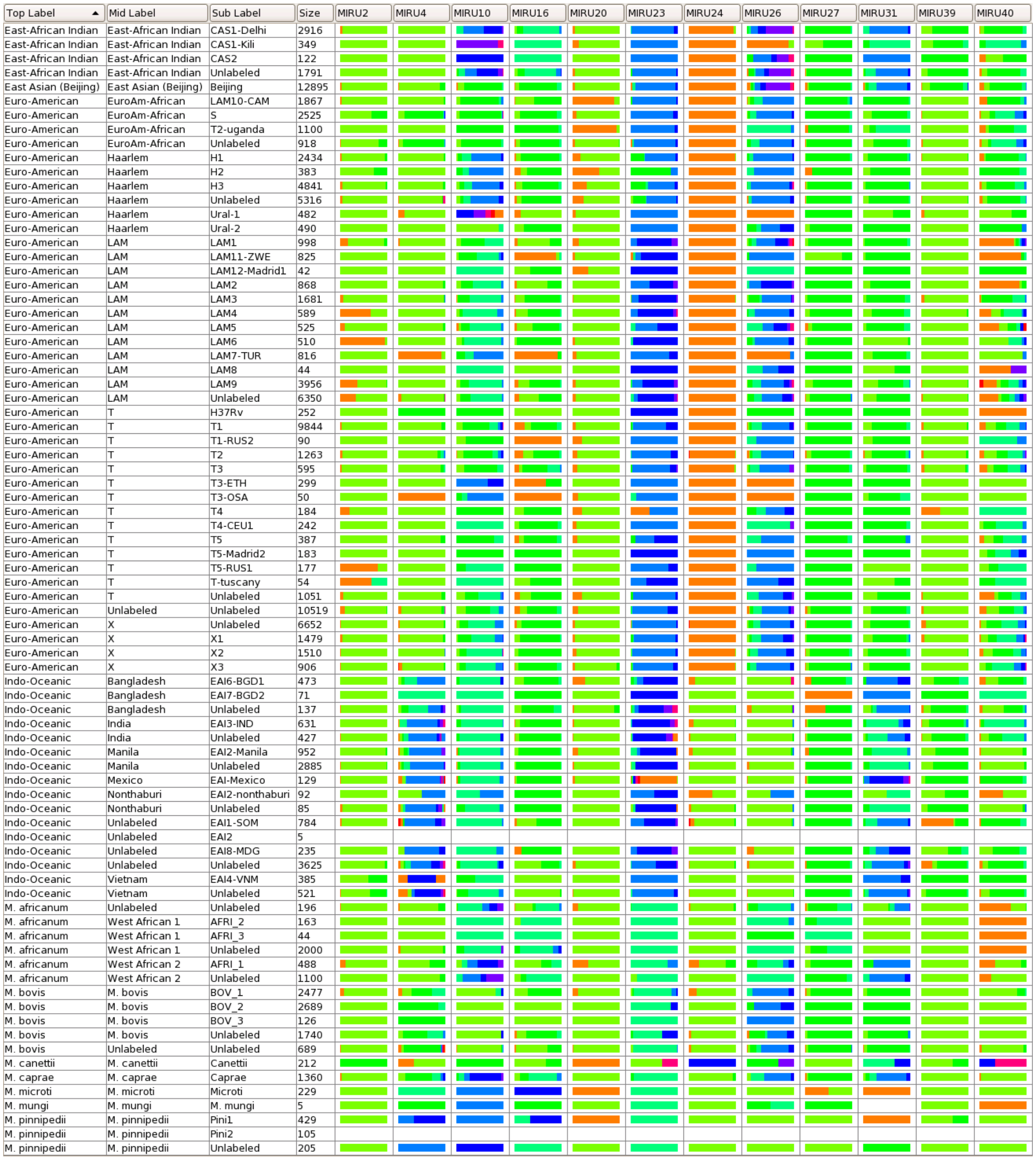
(b) Predicted Labels

Figure 4.10: Original and Predicted Mid-level Data Labels, Final Three-Tier Model. Figure 4.10a shows the original data aggregated to display the mid-level labels, along with mid-level class sizes and average spoligotype / MIRU values. Data which does not possess a mid-level label are listed in this figure as having a mid-level label of "Unlabeled". Figure 4.10b shows the mid-level labels as predicted by the model, with probability, predicted class size (as applied to data set), and spoligtype and MIRU prototypes. The prediction model provides a full label set for every isolate, and thus has no unlabeled data.



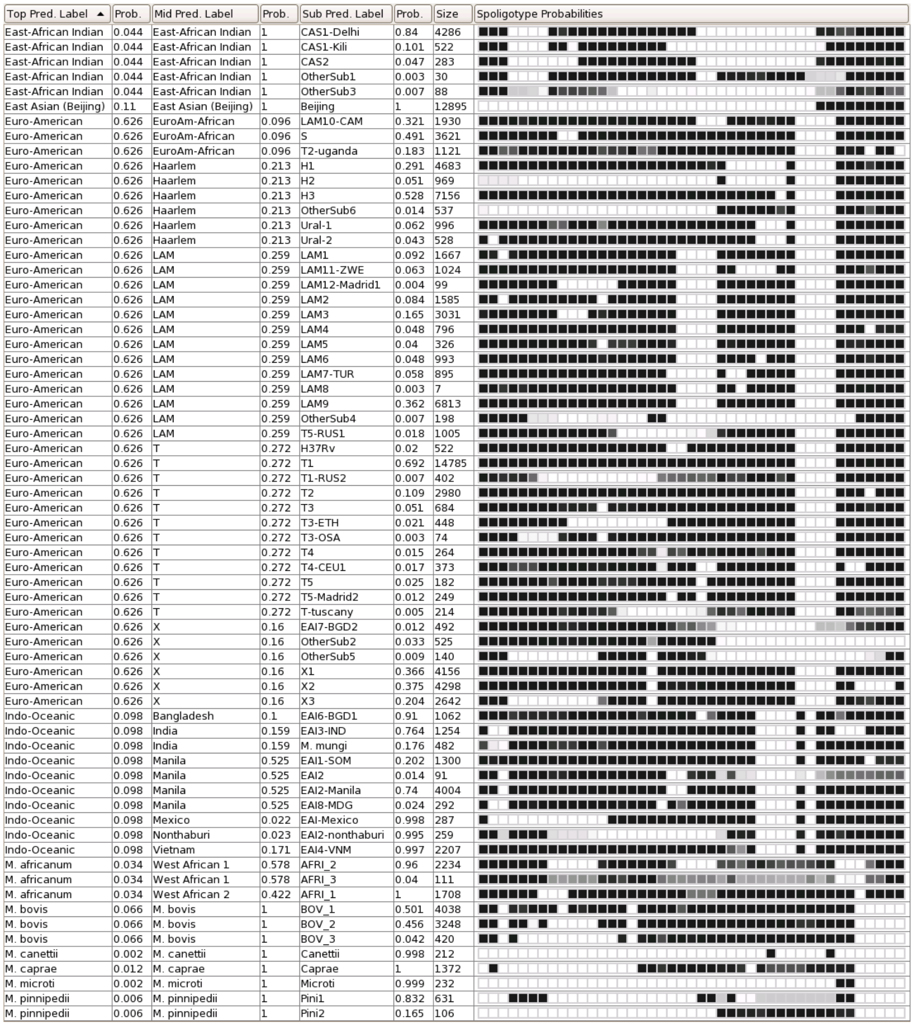
Top Label ▲	Mid Label	Sub Label	Size	Spoligotype Probabilities
East-African Indian	East-African Indian	CAS1-Delhi	2916	
East-African Indian	East-African Indian	CAS1-Kili	349	
East-African Indian	East-African Indian	CAS2	122	
East-African Indian	East-African Indian	Unlabeled	1791	
East Asian (Beijing)	East Asian (Beijing)	Beijing	12895	
Euro-American	EuroAm-African	LAM10-CAM	1867	
Euro-American	EuroAm-African	S	2525	
Euro-American	EuroAm-African	T2-uganda	1100	
Euro-American	EuroAm-African	Unlabeled	918	
Euro-American	Haarlem	H1	2434	
Euro-American	Haarlem	H2	383	
Euro-American	Haarlem	H3	4841	
Euro-American	Haarlem	Unlabeled	5316	
Euro-American	Haarlem	Ural-1	482	
Euro-American	Haarlem	Ural-2	490	
Euro-American	LAM	LAM1	998	
Euro-American	LAM	LAM11-ZWE	825	
Euro-American	LAM	LAM12-Madrid1	42	
Euro-American	LAM	LAM2	868	
Euro-American	LAM	LAM3	1681	
Euro-American	LAM	LAM4	589	
Euro-American	LAM	LAM5	525	
Euro-American	LAM	LAM6	510	
Euro-American	LAM	LAM7-TUR	816	
Euro-American	LAM	LAM8	44	
Euro-American	LAM	LAM9	3956	
Euro-American	LAM	Unlabeled	6350	
Euro-American	T	H37Rv	252	
Euro-American	T	T1	9844	
Euro-American	T	T1-RUS2	90	
Euro-American	T	T2	1263	
Euro-American	T	T3	595	
Euro-American	T	T3-ETH	299	
Euro-American	T	T3-OSA	50	
Euro-American	T	T4	184	
Euro-American	T	T4-CEU1	242	
Euro-American	T	T5	387	
Euro-American	T	T5-Madrid2	183	
Euro-American	T	T5-RUS1	177	
Euro-American	T	T-tuscany	54	
Euro-American	T	Unlabeled	1051	
Euro-American	Unlabeled	Unlabeled	10519	
Euro-American	X	Unlabeled	6652	
Euro-American	X	X1	1479	
Euro-American	X	X2	1510	
Euro-American	X	X3	906	
Indo-Oceanic	Bangladesh	EAI6-BGD1	473	
Indo-Oceanic	Bangladesh	EAI7-BGD2	71	
Indo-Oceanic	Bangladesh	Unlabeled	137	
Indo-Oceanic	India	EAI3-IND	631	
Indo-Oceanic	India	Unlabeled	427	
Indo-Oceanic	Manila	EAI2-Manila	952	
Indo-Oceanic	Manila	Unlabeled	2885	
Indo-Oceanic	Mexico	EAI-Mexico	129	
Indo-Oceanic	Nonthaburi	EAI2-nonthaburi	92	
Indo-Oceanic	Nonthaburi	Unlabeled	85	
Indo-Oceanic	Unlabeled	EAI1-SOM	784	
Indo-Oceanic	Unlabeled	EAI2	5	
Indo-Oceanic	Unlabeled	EAI8-MDG	235	
Indo-Oceanic	Unlabeled	Unlabeled	3625	
Indo-Oceanic	Vietnam	EAI4-VNM	385	
Indo-Oceanic	Vietnam	Unlabeled	521	
M. africanum	Unlabeled	Unlabeled	196	
M. africanum	West African 1	AFRI_2	163	
M. africanum	West African 1	AFRI_3	44	
M. africanum	West African 1	Unlabeled	2000	
M. africanum	West African 2	AFRI_1	488	
M. africanum	West African 2	Unlabeled	1100	
M. bovis	M. bovis	BOV_1	2477	
M. bovis	M. bovis	BOV_2	2689	
M. bovis	M. bovis	BOV_3	126	
M. bovis	M. bovis	Unlabeled	1740	
M. bovis	Unlabeled	Unlabeled	689	
M. canettii	M. canettii	Canettii	212	
M. caprae	M. caprae	Caprae	1360	
M. microti	M. microti	Microti	229	
M. mungi	M. mungi	M. mungi	5	
M. pinnipedii	M. pinnipedii	Pini1	429	
M. pinnipedii	M. pinnipedii	Pini2	105	
M. pinnipedii	M. pinnipedii	Unlabeled	205	

Figure 4.11: Original Spoligotype Data for Sub-level Labels. This figure shows the original data aggregated at the sub-level, displaying the top-level, mid-level, and sub-level label lineages, the size of each lineage in the original data, and the average spoligotype values. Data which does not possess a mid-level or sub-level label are listed in this figure with the label of "Unlabeled". **Figure 4.12 shows the MIRU values for this data set (split across two figures for readability).**



Top Label ▲	Mid Label	Sub Label	Size
East-African Indian	East-African Indian	CAS1-Delhi	2916
East-African Indian	East-African Indian	CAS1-Kili	349
East-African Indian	East-African Indian	CAS2	122
East-African Indian	East-African Indian	Unlabeled	1791
East Asian (Beijing)	East Asian (Beijing)	Beijing	12895
Euro-American	EuroAm-African	LAM10-CAM	1867
Euro-American	EuroAm-African	S	2525
Euro-American	EuroAm-African	T2-uganda	1100
Euro-American	EuroAm-African	Unlabeled	918
Euro-American	Haarlem	H1	2434
Euro-American	Haarlem	H2	383
Euro-American	Haarlem	H3	4841
Euro-American	Haarlem	Unlabeled	5316
Euro-American	Haarlem	Ural-1	482
Euro-American	Haarlem	Ural-2	490
Euro-American	LAM	LAM1	998
Euro-American	LAM	LAM11-ZWE	825
Euro-American	LAM	LAM12-Madrid1	42
Euro-American	LAM	LAM2	868
Euro-American	LAM	LAM3	1681
Euro-American	LAM	LAM4	589
Euro-American	LAM	LAM5	525
Euro-American	LAM	LAM6	510
Euro-American	LAM	LAM7-TUR	816
Euro-American	LAM	LAM8	44
Euro-American	LAM	LAM9	3956
Euro-American	LAM	Unlabeled	6350
Euro-American	T	H37Rv	252
Euro-American	T	T1	9844
Euro-American	T	T1-RUS2	90
Euro-American	T	T2	1263
Euro-American	T	T3	595
Euro-American	T	T3-ETH	299
Euro-American	T	T3-OSA	50
Euro-American	T	T4	184
Euro-American	T	T4-CEU1	242
Euro-American	T	T5	387
Euro-American	T	T5-Madrid2	183
Euro-American	T	T5-RUS1	177
Euro-American	T	T-tuscany	54
Euro-American	T	Unlabeled	1051
Euro-American	Unlabeled	Unlabeled	10519
Euro-American	X	Unlabeled	6652
Euro-American	X	X1	1479
Euro-American	X	X2	1510
Euro-American	X	X3	906
Indo-Oceanic	Bangladesh	EAI6-BGD1	473
Indo-Oceanic	Bangladesh	EAI7-BGD2	71
Indo-Oceanic	Bangladesh	Unlabeled	137
Indo-Oceanic	India	EAI3-IND	631
Indo-Oceanic	India	Unlabeled	427
Indo-Oceanic	Manila	EAI2-Manila	952
Indo-Oceanic	Manila	Unlabeled	2885
Indo-Oceanic	Mexico	EAI-Mexico	129
Indo-Oceanic	Nonthaburi	EAI2-nonthaburi	92
Indo-Oceanic	Nonthaburi	Unlabeled	85
Indo-Oceanic	Unlabeled	EAI1-SOM	784
Indo-Oceanic	Unlabeled	EAI2	5
Indo-Oceanic	Unlabeled	EAI8-MDG	235
Indo-Oceanic	Unlabeled	Unlabeled	3625
Indo-Oceanic	Vietnam	EAI4-VNM	385
Indo-Oceanic	Vietnam	Unlabeled	521
M. africanum	Unlabeled	Unlabeled	196
M. africanum	West African 1	AFRI_2	163
M. africanum	West African 1	AFRI_3	44
M. africanum	West African 1	Unlabeled	2000
M. africanum	West African 2	AFRI_1	488
M. africanum	West African 2	Unlabeled	1100
M. bovis	M. bovis	BOV_1	2477
M. bovis	M. bovis	BOV_2	2689
M. bovis	M. bovis	BOV_3	126
M. bovis	M. bovis	Unlabeled	1740
M. bovis	Unlabeled	Unlabeled	689
M. canettii	M. canettii	Canettii	212
M. caprae	M. caprae	Caprae	1360
M. microti	M. microti	Microti	229
M. mungi	M. mungi	M. mungi	5
M. pinnipedii	M. pinnipedii	Pini1	429
M. pinnipedii	M. pinnipedii	Pini2	105
M. pinnipedii	M. pinnipedii	Unlabeled	205

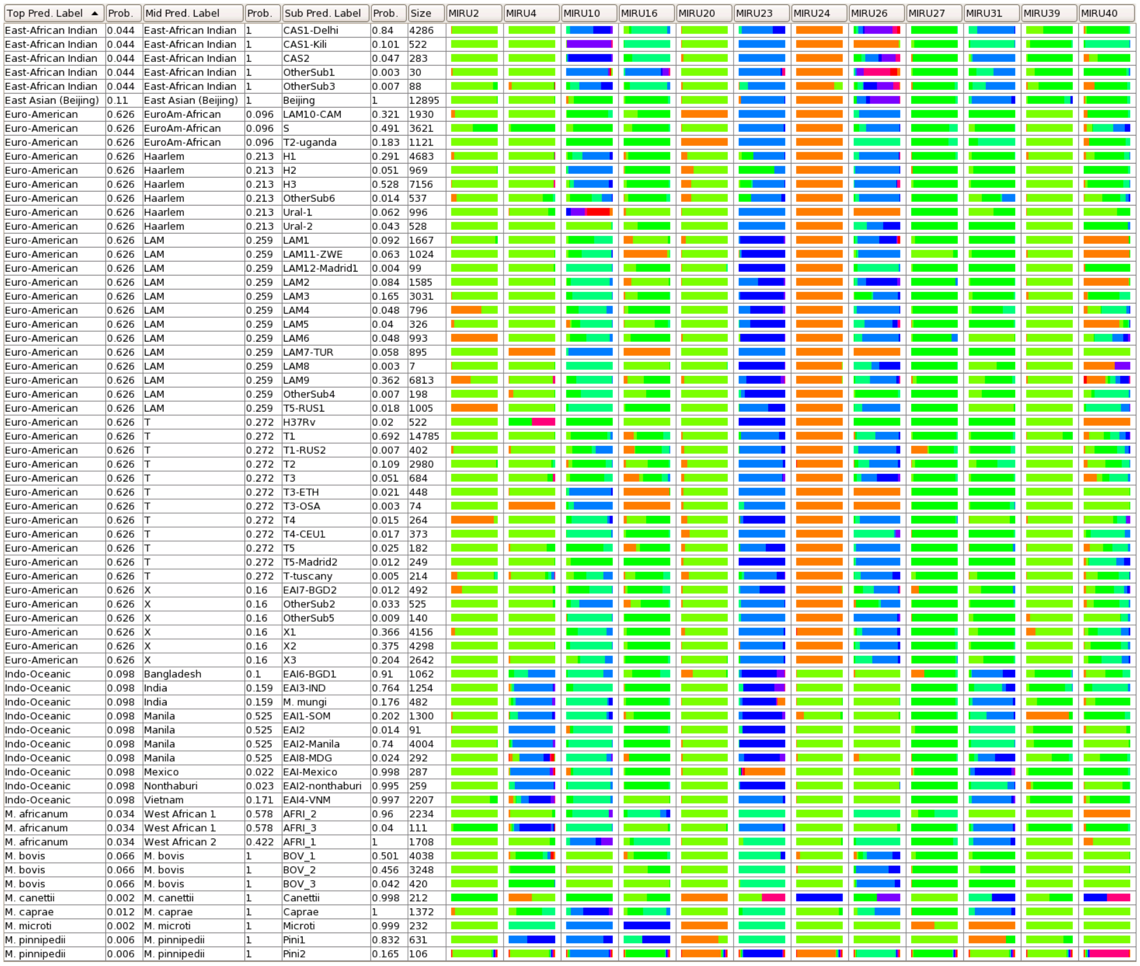
Figure 4.12: Original MIRU Data for Sub-level Labels.
This figure shows the average MIRU values for the same sub-level labels as displayed in Figure 4.11 (split across two figures for readability). The blank MIRU values (for EAI2, *M. Mungi* and Pini2) indicate sub-level classes for which no MIRU data originally existed.



Top Pred. Label ▲	Prob.	Mid Pred. Label	Prob.	Sub Pred. Label	Prob.	Size	Spoligotype Probabilities
East-African Indian	0.044	East-African Indian	1	CAS1-Delhi	0.84	4286	
East-African Indian	0.044	East-African Indian	1	CAS1-Kili	0.101	522	
East-African Indian	0.044	East-African Indian	1	CAS2	0.047	283	
East-African Indian	0.044	East-African Indian	1	OtherSub1	0.003	30	
East-African Indian	0.044	East-African Indian	1	OtherSub3	0.007	88	
East Asian (Beijing)	0.11	East Asian (Beijing)	1	Beijing	1	12895	
Euro-American	0.626	EuroAm-African	0.096	LAM10-CAM	0.321	1930	
Euro-American	0.626	EuroAm-African	0.096	S	0.491	3621	
Euro-American	0.626	EuroAm-African	0.096	T2-uganda	0.183	1121	
Euro-American	0.626	Haarlem	0.213	H1	0.291	4683	
Euro-American	0.626	Haarlem	0.213	H2	0.051	969	
Euro-American	0.626	Haarlem	0.213	H3	0.528	7156	
Euro-American	0.626	Haarlem	0.213	OtherSub6	0.014	537	
Euro-American	0.626	Haarlem	0.213	Ural-1	0.062	996	
Euro-American	0.626	Haarlem	0.213	Ural-2	0.043	528	
Euro-American	0.626	LAM	0.259	LAM1	0.092	1667	
Euro-American	0.626	LAM	0.259	LAM11-ZWE	0.063	1024	
Euro-American	0.626	LAM	0.259	LAM12-Madrid1	0.004	99	
Euro-American	0.626	LAM	0.259	LAM2	0.084	1585	
Euro-American	0.626	LAM	0.259	LAM3	0.165	3031	
Euro-American	0.626	LAM	0.259	LAM4	0.048	796	
Euro-American	0.626	LAM	0.259	LAM5	0.04	326	
Euro-American	0.626	LAM	0.259	LAM6	0.048	993	
Euro-American	0.626	LAM	0.259	LAM7-TUR	0.058	895	
Euro-American	0.626	LAM	0.259	LAM8	0.003	7	
Euro-American	0.626	LAM	0.259	LAM9	0.362	6813	
Euro-American	0.626	LAM	0.259	OtherSub4	0.007	198	
Euro-American	0.626	LAM	0.259	T5-RUS1	0.018	1005	
Euro-American	0.626	T	0.272	H37Rv	0.02	522	
Euro-American	0.626	T	0.272	T1	0.692	14785	
Euro-American	0.626	T	0.272	T1-RUS2	0.007	402	
Euro-American	0.626	T	0.272	T2	0.109	2980	
Euro-American	0.626	T	0.272	T3	0.051	684	
Euro-American	0.626	T	0.272	T3-ETH	0.021	448	
Euro-American	0.626	T	0.272	T3-OSA	0.003	74	
Euro-American	0.626	T	0.272	T4	0.015	264	
Euro-American	0.626	T	0.272	T4-CEU1	0.017	373	
Euro-American	0.626	T	0.272	T5	0.025	182	
Euro-American	0.626	T	0.272	T5-Madrid2	0.012	249	
Euro-American	0.626	T	0.272	T-tuscany	0.005	214	
Euro-American	0.626	X	0.16	EAI7-BGD2	0.012	492	
Euro-American	0.626	X	0.16	OtherSub2	0.033	525	
Euro-American	0.626	X	0.16	OtherSub5	0.009	140	
Euro-American	0.626	X	0.16	X1	0.366	4156	
Euro-American	0.626	X	0.16	X2	0.375	4298	
Euro-American	0.626	X	0.16	X3	0.204	2642	
Indo-Oceanic	0.098	Bangladesh	0.1	EAI6-BGD1	0.91	1062	
Indo-Oceanic	0.098	India	0.159	EAI3-IND	0.764	1254	
Indo-Oceanic	0.098	India	0.159	M. mungi	0.176	482	
Indo-Oceanic	0.098	Manila	0.525	EAI1-SOM	0.202	1300	
Indo-Oceanic	0.098	Manila	0.525	EAI2	0.014	91	
Indo-Oceanic	0.098	Manila	0.525	EAI2-Manila	0.74	4004	
Indo-Oceanic	0.098	Manila	0.525	EAI8-MDG	0.024	292	
Indo-Oceanic	0.098	Mexico	0.022	EAI-Mexico	0.998	287	
Indo-Oceanic	0.098	Nonthaburi	0.023	EAI2-nonthaburi	0.995	259	
Indo-Oceanic	0.098	Vietnam	0.171	EAI4-VNM	0.997	2207	
M. africanum	0.034	West African 1	0.578	AFRI_2	0.96	2234	
M. africanum	0.034	West African 1	0.578	AFRI_3	0.04	111	
M. africanum	0.034	West African 2	0.422	AFRI_1	1	1708	
M. bovis	0.066	M. bovis	1	BOV_1	0.501	4038	
M. bovis	0.066	M. bovis	1	BOV_2	0.456	3248	
M. bovis	0.066	M. bovis	1	BOV_3	0.042	420	
M. canettii	0.002	M. canettii	1	Canettii	0.998	212	
M. caprae	0.012	M. caprae	1	Caprae	1	1372	
M. microti	0.002	M. microti	1	Microti	0.999	232	
M. pinnipedii	0.006	M. pinnipedii	1	Pini1	0.832	631	
M. pinnipedii	0.006	M. pinnipedii	1	Pini2	0.165	106	

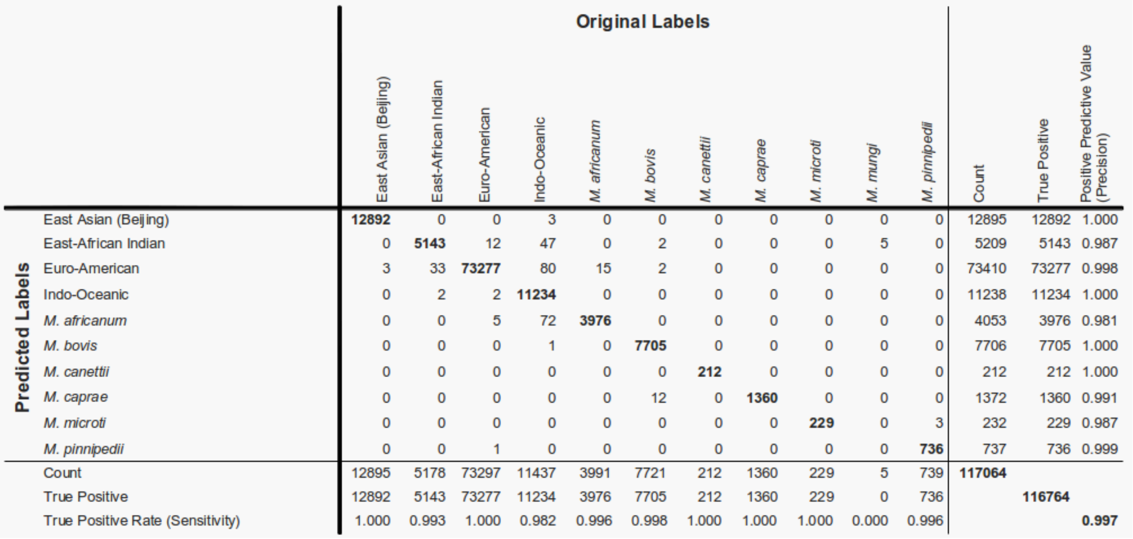
Figure 4.13: Predicted Spoligotype Model for Sub-level Labels, Final Three-Tier Model.

This figure shows the sub-level labels as predicted by the model, with probability, predicted class size (as applied to data set), and spoligtype prototypes. The prediction model provides a full label set for every isolate, and thus has no unlabeled data. This figure can be compared to Figure 4.11 to examine differences between the data as originally labeled and as labeled by the model.



Top Pred. Label ▲	Prob.	Mid Pred. Label	Prob.	Sub Pred. Label	Prob.	Size	MIRU2	MIRU4	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40
East-African Indian	0.044	East-African Indian	1	CAS1-Delhi	0.84	4286												
East-African Indian	0.044	East-African Indian	1	CAS1-Kili	0.101	522												
East-African Indian	0.044	East-African Indian	1	CAS2	0.047	283												
East-African Indian	0.044	East-African Indian	1	OtherSub1	0.003	30												
East-African Indian	0.044	East-African Indian	1	OtherSub3	0.007	88												
East Asian (Beijing)	0.11	East Asian (Beijing)	1	Beijing	1	12895												
Euro-American	0.626	EuroAm-African	0.096	LAM10-CAM	0.321	1930												
Euro-American	0.626	EuroAm-African	0.096	S	0.491	3621												
Euro-American	0.626	EuroAm-African	0.096	T2-uganda	0.183	1121												
Euro-American	0.626	Haarlem	0.213	H1	0.291	4683												
Euro-American	0.626	Haarlem	0.213	H2	0.051	969												
Euro-American	0.626	Haarlem	0.213	H3	0.528	7156												
Euro-American	0.626	Haarlem	0.213	OtherSub6	0.014	537												
Euro-American	0.626	Haarlem	0.213	Ural-1	0.062	996												
Euro-American	0.626	Haarlem	0.213	Ural-2	0.043	528												
Euro-American	0.626	LAM	0.259	LAM1	0.092	1667												
Euro-American	0.626	LAM	0.259	LAM11-ZWE	0.063	1024												
Euro-American	0.626	LAM	0.259	LAM12-Madrid1	0.004	99												
Euro-American	0.626	LAM	0.259	LAM2	0.084	1585												
Euro-American	0.626	LAM	0.259	LAM3	0.165	3031												
Euro-American	0.626	LAM	0.259	LAM4	0.048	796												
Euro-American	0.626	LAM	0.259	LAM5	0.04	326												
Euro-American	0.626	LAM	0.259	LAM6	0.048	993												
Euro-American	0.626	LAM	0.259	LAM7-TUR	0.058	895												
Euro-American	0.626	LAM	0.259	LAM8	0.003	7												
Euro-American	0.626	LAM	0.259	LAM9	0.362	6813												
Euro-American	0.626	LAM	0.259	OtherSub4	0.007	198												
Euro-American	0.626	LAM	0.259	T5-RUS1	0.018	1005												
Euro-American	0.626	T	0.272	H37Rv	0.02	522												
Euro-American	0.626	T	0.272	T1	0.692	14785												
Euro-American	0.626	T	0.272	T1-RUS2	0.007	402												
Euro-American	0.626	T	0.272	T2	0.109	2980												
Euro-American	0.626	T	0.272	T3	0.051	684												
Euro-American	0.626	T	0.272	T3-ETH	0.021	448												
Euro-American	0.626	T	0.272	T3-OSA	0.003	74												
Euro-American	0.626	T	0.272	T4	0.015	264												
Euro-American	0.626	T	0.272	T4-CEU1	0.017	373												
Euro-American	0.626	T	0.272	T5	0.025	182												
Euro-American	0.626	T	0.272	T5-Madrid2	0.012	249												
Euro-American	0.626	T	0.272	T-tuscany	0.005	214												
Euro-American	0.626	X	0.16	EAI7-BGD2	0.012	492												
Euro-American	0.626	X	0.16	OtherSub2	0.033	525												
Euro-American	0.626	X	0.16	OtherSub5	0.009	140												
Euro-American	0.626	X	0.16	X1	0.366	4156												
Euro-American	0.626	X	0.16	X2	0.375	4298												
Euro-American	0.626	X	0.16	X3	0.204	2642												
Indo-Oceanic	0.098	Bangladesh	0.1	EAI6-BGD1	0.91	1062												
Indo-Oceanic	0.098	India	0.159	EAI3-IND	0.764	1254												
Indo-Oceanic	0.098	India	0.159	M. mungi	0.176	482												
Indo-Oceanic	0.098	Manila	0.525	EAI1-SOM	0.202	1300												
Indo-Oceanic	0.098	Manila	0.525	EAI2	0.014	91												
Indo-Oceanic	0.098	Manila	0.525	EAI2-Manila	0.74	4004												
Indo-Oceanic	0.098	Manila	0.525	EAI8-MDG	0.024	292												
Indo-Oceanic	0.098	Mexico	0.022	EAI-Mexico	0.998	287												
Indo-Oceanic	0.098	Nonthaburi	0.023	EAI2-nonthaburi	0.995	259												
Indo-Oceanic	0.098	Vietnam	0.171	EAI4-VNM	0.997	2207												
M. africanum	0.034	West African 1	0.578	AFRI_2	0.96	2234												
M. africanum	0.034	West African 1	0.578	AFRI_3	0.04	111												
M. africanum	0.034	West African 2	0.422	AFRI_1	1	1708												
M. bovis	0.066	M. bovis	1	BOV_1	0.501	4038												
M. bovis	0.066	M. bovis	1	BOV_2	0.456	3248												
M. bovis	0.066	M. bovis	1	BOV_3	0.042	420												
M. canettii	0.002	M. canettii	1	Canettii	0.998	212												
M. caprae	0.012	M. caprae	1	Caprae	1	1372												
M. microti	0.002	M. microti	1	Microti	0.999	232												
M. pinnipedii	0.006	M. pinnipedii	1	Pini1	0.832	631												
M. pinnipedii	0.006	M. pinnipedii	1	Pini2	0.165	106												

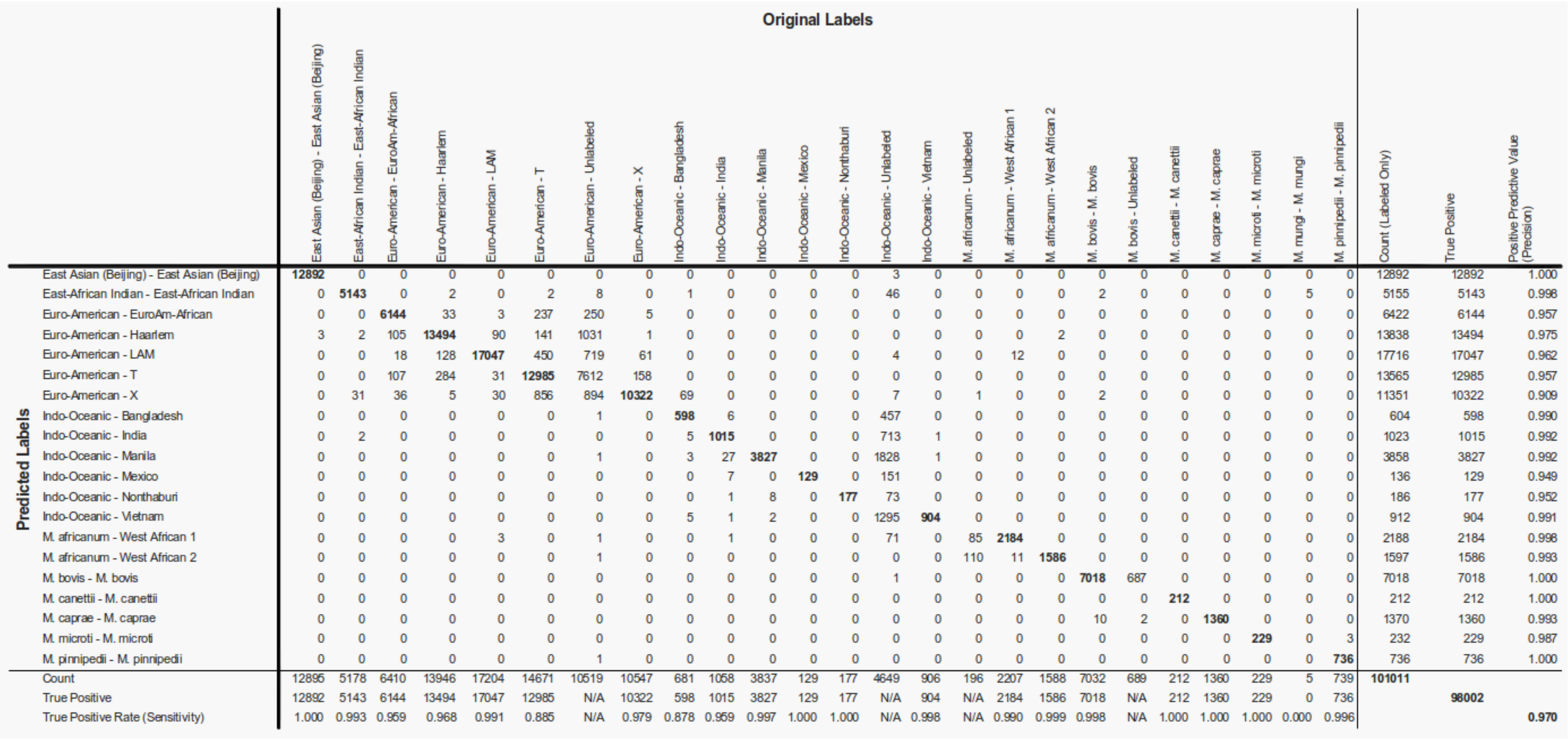
Figure 4.14: Predicted MIRU Model for Sub-level Labels, Final Three-Tier Model.
This figure shows the average MIRU values for the same predicted sub-level labels as displayed in Figure 4.13 (split across two figures for readability). This figure can be compared to Figure 4.12 to examine differences between the data as originally labeled and as labeled by the model.



		Original Labels													
		East Asian (Beijing)	East-African Indian	Euro-American	Indo-Oceanic	*M. africanum*	*M. bovis*	*M. canettii*	*M. caprae*	*M. microti*	*M. mungi*	*M. pinnipedii*	Count	True Positive	Positive Predictive Value (Precision)
Predicted Labels	East Asian (Beijing)	**12892**	0	0	3	0	0	0	0	0	0	0	12895	12892	1.000
	East-African Indian	0	**5143**	12	47	0	2	0	0	0	5	0	5209	5143	0.987
	Euro-American	3	33	**73277**	80	15	2	0	0	0	0	0	73410	73277	0.998
	Indo-Oceanic	0	2	2	**11234**	0	0	0	0	0	0	0	11238	11234	1.000
	M. africanum	0	0	5	72	**3976**	0	0	0	0	0	0	4053	3976	0.981
	M. bovis	0	0	0	1	0	**7705**	0	0	0	0	0	7706	7705	1.000
	M. canettii	0	0	0	0	0	0	**212**	0	0	0	0	212	212	1.000
	M. caprae	0	0	0	0	0	12	0	**1360**	0	0	0	1372	1360	0.991
	M. microti	0	0	0	0	0	0	0	0	**229**	0	3	232	229	0.987
	M. pinnipedii	0	0	1	0	0	0	0	0	0	0	**736**	737	736	0.999
	Count	12895	5178	73297	11437	3991	7721	212	1360	229	5	739	**117064**		
	True Positive	12892	5143	73277	11234	3976	7705	212	1360	229	0	736		**116764**	
	True Positive Rate (Sensitivity)	1.000	0.993	1.000	0.982	0.996	0.998	1.000	1.000	1.000	0.000	0.996			**0.997**

Figure 4.15: Confusion Matrix for Original and Predicted Top-Level Labels, Final Three-Tier Model.

Shows a comparison of the original data set and the data set at the top-most level, as predicted by the model by displaying the number of correctly (in bold) and incorrectly labeled isolates. For each original top-level label, the total count, number of correctly-predicted isolates (true positives), and the model sensitivity (number of true positives divided by the count per original label) are displayed at the bottom. For each predicted label, the total count, number of correctly-predicated isolates, and model precision (number of true positives divided by the count per *predicted* label) are displayed at the right. The totals in the bottom right square represent (from first to last) the total count in the database, the number of correctly-predicted labels, and accuracy of the model at the top-most level.



		Original Labels																										
		East Asian (Beijing) - East Asian (Beijing)	East-African Indian - East-African Indian	Euro-American - EuroAm-African	Euro-American - Haarlem	Euro-American - LAM	Euro-American - T	Euro-American - Unlabeled	Euro-American - X	Indo-Oceanic - Bangladesh	Indo-Oceanic - India	Indo-Oceanic - Manila	Indo-Oceanic - Mexico	Indo-Oceanic - Nonthaburi	Indo-Oceanic - Unlabeled	Indo-Oceanic - Vietnam	M. africanum - Unlabeled	M. africanum - West African 1	M. africanum - West African 2	M. bovis - M. bovis	M. bovis - Unlabeled	M. canettii - M. canettii	M. caprae - M. caprae	M. microti - M. microti	M. mungi - M. mungi	M. pinnipedii - M. pinnipedii	Count (Labeled Only)	True Positive	Positive Predictive Value (Precision)
Predicted Labels	East Asian (Beijing) - East Asian (Beijing)	**12892**	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	12892	12892	1.000
	East-African Indian - East-African Indian	0	**5143**	0	2	0	2	8	0	1	0	0	0	0	46	0	0	0	0	2	0	0	0	0	5	0	5155	5143	0.998
	Euro-American - EuroAm-African	0	0	**6144**	33	3	237	250	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6422	6144	0.957
	Euro-American - Haarlem	3	2	105	**13494**	90	141	1031	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	13838	13494	0.975
	Euro-American - LAM	0	0	18	128	**17047**	450	719	61	0	0	0	0	0	4	0	0	12	0	0	0	0	0	0	0	0	17716	17047	0.962
	Euro-American - T	0	0	107	284	31	**12985**	7612	158	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13565	12985	0.957
	Euro-American - X	0	31	36	5	30	856	894	**10322**	69	0	0	0	0	7	0	1	0	0	2	0	0	0	0	0	0	11351	10322	0.909
	Indo-Oceanic - Bangladesh	0	0	0	0	0	0	1	0	**598**	6	0	0	0	457	0	0	0	0	0	0	0	0	0	0	0	604	598	0.990
	Indo-Oceanic - India	0	2	0	0	0	0	0	0	5	**1015**	0	0	0	713	1	0	0	0	0	0	0	0	0	0	0	1023	1015	0.992
	Indo-Oceanic - Manila	0	0	0	0	0	0	1	0	3	27	**3827**	0	0	1828	1	0	0	0	0	0	0	0	0	0	0	3858	3827	0.992
	Indo-Oceanic - Mexico	0	0	0	0	0	0	0	0	0	7	0	**129**	0	151	0	0	0	0	0	0	0	0	0	0	0	136	129	0.949
	Indo-Oceanic - Nonthaburi	0	0	0	0	0	0	0	0	0	1	8	0	**177**	73	0	0	0	0	0	0	0	0	0	0	0	186	177	0.952
	Indo-Oceanic - Vietnam	0	0	0	0	0	0	0	0	5	1	2	0	0	1295	**904**	0	0	0	0	0	0	0	0	0	0	912	904	0.991
	M. africanum - West African 1	0	0	0	0	3	0	1	0	0	1	0	0	0	71	0	85	**2184**	0	0	0	0	0	0	0	0	2188	2184	0.998
	M. africanum - West African 2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	110	11	**1586**	0	0	0	0	0	0	0	1597	1586	0.993
	M. bovis - M. bovis	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	**7018**	687	0	0	0	0	0	7018	7018	1.000
	M. canettii - M. canettii	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**212**	0	0	0	0	212	212	1.000
	M. caprae - M. caprae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	2	0	**1360**	0	0	0	1370	1360	0.993
	M. microti - M. microti	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**229**	0	3	232	229	0.987
	M. pinnipedii - M. pinnipedii	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**736**	736	736	1.000
	Count	12895	5178	6410	13946	17204	14671	10519	10547	681	1058	3837	129	177	4649	906	196	2207	1588	7032	689	212	1360	229	5	739	**101011**		
	True Positive	12892	5143	6144	13494	17047	12985	N/A	10322	598	1015	3827	129	177	N/A	904	N/A	2184	1586	7018	N/A	212	1360	229	0	736		**98002**	
	True Positive Rate (Sensitivity)	1.000	0.993	0.959	0.968	0.991	0.885	N/A	0.979	0.878	0.959	0.997	1.000	1.000	N/A	0.998	N/A	0.990	0.999	0.998	N/A	1.000	1.000	1.000	0.000	0.996			**0.970**

Figure 4.16: Confusion Matrix for Original and Predicted Mid-Level Labels, Final Three-Tier Model. Shows a comparison of the original data set and the data set at the mid-level, predicted by the model by displaying the number of correctly (in bold) and incorrectly labeled isolates. For each original mid-level label, the total count, number of correctly-predicted isolates (true positives), and the model sensitivity (number of true positives divided by the count per original label) are displayed at the bottom. For each predicted label, the count (excluding the isolates that were originally unlabeled), number of correctly-predicated isolates, and model precision (number of true positives divided by the count per *predicted* label) are displayed at the right. The totals in the bottom right square represent (from first to last) the total count of *labeled* in the database, the number of correctly-predicted labels, and accuracy of the model at the mid-level. Note that the total number of isolates in this table is the same as in Figure 4.15; the total listed in the bottom-right only shows the count of the fully-labeled data, since that count is important for the calculation of the model precision and accuracy.

CHAPTER 5
Conclusion and Future Work

This thesis presents a new probabilistic model that predicts all three levels of the MTBC lineage hierarchy based on spoligotyping and MIRU data simultaneously. This model handles MTBC isolates with spoligotypes and MIRU as well as with spoligotypes only, and handles partially-labeled data, which are missing one or more of the top-level, mid-level, or sub-level labels due to heterogeneous data sources. This work uses a semi-supervised method for lineage prediction which builds upon previous unsupervised and supervised methods for single-level lineage prediction.

The results of this thesis confirm several well-established and well-studied sublineages as well as identify several possible new sublineages which may require additional study. Additionally, some possibly inconsistent lineage labels have been identified, specifically the EAI7-BGD2 sublineage (as identified by Institut Pasteur), which is considered to be a Euro-American/X sublineage after grouping it with similar isolates provided by CDC.

Certain assumptions and limitations of the three-tier model presented in this work could be improved upon in future efforts. The failure to properly classify the *M. Mungi* lineage can be rectified by upsampling the *M. mungi* isolates to increase their sample size, since this lineage is very rare. The independence assumption for spoligotype spacers does not hold true in real-world data, but nonetheless has been shown to provide effective classification results. However, more complex models for spoligotype spacers may provide even better classification results. Additionally, modeling MIRU data as a categorical distribution provides simplicity and ease of computation, but as MIRU values are ordinal in nature, the categorical distribution ignores the “closeness” of two values. In some cases, such at MIRU24, this proves useful, as the difference between a MIRU24 value of 1 and a MIRU24 value of 2 is an important predictor. However, in other cases, modeling MIRU values as a categorical distribution may be introducing unnecessary separation between two otherwise closely related isolates. A model which considers the ordinality of MIRU values may prove to be a better classifier than the one presented in this thesis.

Future work to be conducted in this study includes resolving the EAI7-BGD2 issue through a literature search and consultation with Institut Pasteur and CDC. The model will be retrained based on successful resolution of problem classifying *M. mungi* isolates through upsampling, and further evaluation of this model will be done on an additional data set provided by Nalin Rastogi and David Couvin to generate additional out-of-sample predictions.

Following the precedent set by SPOTCLUST [11] and TB-Lineage [13], once completed, the final model will be made available through the TB-Insight project website (`http://tbinsight.cs.rpi.edu/`) and SITVITWEB, for the benefit of TB public health workers and MTBC resarchers worldwide.

REFERENCES

[1] Centers for Disease Control and Prevention, Atlanta, GA, *Guide to the Application of Genotyping to Tuberculosis Prevention and Control*, 2004.

[2] C. Demay, B. Liens, T. Burguiere, V. Hill, D. Couvin, J. Millet, I. Mokrousov, C. Sola, T. Zozio, and N. Rastogi, "SITVITWEB–a publicly available international multimarker database for studying mycobacterium tuberculosis genetic diversity and molecular epidemiology," *Infect. Genet. Evol.*, vol. 12, no. 4, pp. 755–66, 2012.

[3] C. Allix-Beguec, D. Harmsen, T. Weniger, P. Supply, and S. Neimann, "Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of mycobacterium tuberculosis complex isolates," *J. Clin. Microbiol.*, vol. 46, no. 8, pp. 2692–9, 2008.

[4] T. Weniger, J. Krawczyk, P. Supply, S. Niemann, and D. Harmsen, "MIRU-VNTRplus: a web tool for polyphasic genotyping of mycobacterium tuberculosis complex bacteria," *Nucleic Acids Res.*, no. 38, pp. 326–331, 2010.

[5] 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.*, vol. 35, no. 4, pp. 907–14, 1997.

[6] P. Supply, C. Allix, S. Lesjean, M. Cardoso-Oelemann, S. Rüsch-Gerdes, E. Willery, E. Savine, P. de Haas, H. van Deutekom, S. Roring, P. Bifani, N. Kurepina, B. Kreiswirth, C. Sola, N. Rastogi, V. Vatin, M. Gutierrez, M. Fauville, S. Neimann, R. Skuce, K. Kremer, C. Locht, and D. van Soolingen, "Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of

mycobacterium tuberculosis," *J. Clin. Microbiol.*, vol. 44, no. 12, pp. 4498–510, 2006.

[7] S. Ferdinand, G. Valétudie, 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.*, vol. 155, no. 8, pp. 647–654, 2004.

[8] I. Filliol, J. R. Driscoll, D. van Soolingen, B. N. Kreiswirth, K. Kremer, G. Valétudie, D. D. Anh, R. Barlow, D. Banerjee, P. J. Bifani, *et al.*, "Snapshot of moving and expanding clones of mycobacterium tuberculosis and their global distribution assessed by spoligotyping in an international study," *J. Clin. Microbiol.*, vol. 41, no. 5, pp. 1963–1970, 2003.

[9] T. Wirth, F. Hildebrand, C. Allix-Béguec, F. Wölbeling, T. Kubica, K. Kremer, D. van Soolingen, S. Rüsch-Gerdes, C. Locht, S. Brisse, A. Meyer, P. Supply, and S. Niemann, "Origin, spread and demography of the mycobacterium tuberculosis complex," *PLoS Pathog.*, vol. 4, no. 9, p. e1000160, 2008.

[10] K. Brudey, J. R. Driscoll, L. Rigouts, W. M. Prodinger, A. Gori, S. A. Al-Hajoj, C. Allix, L. Aristimuño, J. Arora, V. Baumanis, *et al.*, "Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology," *BMC Microbiol.*, vol. 6, no. 1, p. 23, 2006.

[11] I. Vitol, "Mathematical models for mycobacterium tuberculosis complex genotyping and patient data," Ph.D. dissertation, Dept. Comput. Sci., Rensselaer Polytechnic Inst., Troy, NY, 2006.

[12] M. Aminian, A. Shabbeer, and K. P. Bennett, "A conformal bayesian network for classification of mycobacterium tuberculosis complex lineages," *BMC Bioinformatics*, vol. 11, no. Suppl. 3, p. S4, 2010.

[13] A. Shabbeer, L. S. Cowan, C. Ozcaglar, N. Rastogi, S. L. Vandenberg, B. Yener, and K. P. Bennett, "TB-Lineage: an online tool for classification and analysis of strains of mycobacterium tuberculosis complex," *Infect. Genet. Evol.*, vol. 12, no. 4, pp. 789–97, 2012.

[14] Y.-J. Sun, A. S. G. Lee, S. T. Ng, S. Ravindran, K. Kremer, R. Bellamy, S.-Y. Wong, D. van Soolingen, P. Supply, and N. I. Paton, "Characterization of ancestral mycobacterium tuberculosis by multiple genetic markers and proposal of genotyping strategy," *J. Clin. Microbiol.*, vol. 42, no. 11, pp. 5058–5064, 2004.

[15] I. Comas, S. Homolka, S. Niemann, and S. Gagneux, "Genotyping of genetically monomorphic bacteria: DNA sequencing in mycobacterium tuberculosis highlights the limitations of current methodologies," *PLoS ONE*, vol. 11, no. 4, p. e7815, 2009.

[16] I. Mokrousov, "The quiet and controversial: Ural family of mycobacterium tuberculosis," *Infect. Genet. Evol.*, vol. 12, no. 4, pp. 619–29, 2012.

[17] J. D. 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.*, vol. 182, no. 9, pp. 2393–401, 2000.

[18] R. Warren, E. Streicher, S. Sampson, G. van der Spuy, M. Richardson, D. Nguyen, M. Behr, T. Victor, and P. van Helden, "Microevolution of the direct repeat region of mycobacterium tuberculosis: implications for interpretation of spoligotyping data," *J. Clin. Microbiol.*, vol. 40, no. 12, pp. 4457–65, 2002.

[19] 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.*, vol. 42, no. 11, pp. 5388–91, 2004.

[20] Z. Fang, N. Morrison, B. Watt, and K. J. Forbes, “IS6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related mycobacterium tuberculosis strains,” *J. Bacteriol.*, vol. 180, no. 8, pp. 2102–2109, 1998.

[21] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: MIT Press, 2012.

[22] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, NY: Springer, 2nd ed., 2009.

[23] K. Nigam, A. Mccallum, and T. Mitchell, “Semi-supervised text classification using EM,” in *Semi-Supervised Learning* (O. Chapelle, A. Zien, and B. Schölkopf, eds.), Cambridge, MA: MIT Press, 2006.

[24] P. Smyth, “Model selection for probabilistic clustering using cross-validated likelihood,” *Stat. Comput.*, vol. 10, no. 1, pp. 63–72, 2000.

[25] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716—-723, 1974.

[26] G. Schwarz, “Estimating the dimension of a model,” *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

APPENDIX A
EM Algorithm Update Derivations

This appendix describes the derivation of the expectation-maximization algorithm update steps.

A.1 Single-tier Sublineage Model Updates

This section describes in detail the derivations required for the expectation-maximization algorithm as applied to the single-tier model without hidden parrents, originally described in Section 3.1.1.

A.1.1 Maximization for $\alpha_{\hat{j}}$

Optimization with respect to a particular $\alpha_{\hat{j}}$ is constrained by $\sum_{j=1}^{k} \alpha_j = 1$. Thus, the optimality condition with respect to the parameter $\alpha_{\hat{j}}$ of Equation 3.1 is

$$\frac{\partial}{\partial \alpha_{\hat{j}}} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} E_Z \left[z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] \log \left\{ p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} + \lambda \left(\sum_{j=1}^{k} \alpha_j - 1 \right) \right] = 0$$

Noting that

$$p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) = p(c_j) p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) = \alpha_j p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) \ ,$$

we continue the above computation as

$$\begin{aligned} \frac{\partial}{\partial \alpha_{\hat{j}}} \Bigg[\sum_{i=1}^{n} \sum_{j=1}^{k} \big(& E_Z \left[z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] \log \left\{ \alpha_j \right\} + \\ & E_Z \left[z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] \log \left\{ p(\mathbf{s}_i \mid c_j) p(\mathbf{m}_i \mid c_j) \right\} \big) + \lambda \left(\sum_{j=1}^{k} \alpha_j - 1 \right) \Bigg] = 0 \ , \end{aligned}$$

which simplifies as

$$\frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{\alpha_{\hat{j}}} + \lambda = 0$$

or equivalently

$$\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] + \alpha_{\hat{j}}\lambda = 0 \ .$$

To compute the Lagrange multiplier λ, we sum up over all $\partial/\partial\alpha_{\hat{j}}$:

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{i=1}^{n} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] + \sum_{j=1}^{k}\alpha_j\lambda = 0 & \\
\lambda &= -\sum_{j=1}^{k}\sum_{i=1}^{n} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] \\
\lambda &= -n
\end{aligned}$$

where the calculation follows as we note that $\sum_{j=1}^{k}\alpha_j = 1$ (by constraint) and for a particular i,

$$\sum_{j=1}^{k} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] = \sum_{j=1}^{k} p(c_j \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}) = 1 \ .$$

Thus,

$$\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] - \alpha_{\hat{j}}n = 0$$

which gives us our $\alpha_{\hat{j}}$ estimate update

$$\alpha_{\hat{j}} = \frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{n} \ .$$

A.1.2 Maximization for $\sigma_{\hat{j}\hat{d}}$

The optimality condition of Equation 3.1 with respect to the parameter $\sigma_{\hat{j}\hat{d}}$ is

$$\frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \log\left\{p(\mathbf{s}_i, \mathbf{m}_i \mid \boldsymbol{\Theta})\right\}\right] = 0$$

Noting that

$$\begin{aligned}\log\{p(\mathbf{s}_i, \mathbf{m}_i \mid \boldsymbol{\Theta})\} &= \log\{p(c_j)p(\mathbf{s}_i \mid c_j)p(\mathbf{m}_i \mid c_j)\}\\ &= \log\{p(c_j)\} + \log\{p(\mathbf{s}_i \mid c_j)\} + \log\{p(\mathbf{m}_i \mid c_j)\}\end{aligned}$$

and that $\log\{p(\mathbf{s}_i \mid c_j)\}$ term is the only term with contains the parameter $\sigma_{\hat{j}\hat{d}}$, we can simplify this computation as

$$\begin{aligned}\frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \log\{p(\mathbf{s}_i \mid c_j)\}\right] &= 0\\ \frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \log\left\{\prod_{d=1}^{D_S} \sigma_{jd}^{s_{id}}(1-\sigma_{jd})^{1-s_{id}}\right\}\right] &= 0\\ \sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\log\left\{\sigma_{\hat{j}\hat{d}}^{s_{i\hat{d}}}(1-\sigma_{\hat{j}\hat{d}})^{1-s_{i\hat{d}}}\right\}\right] &= 0\ .\end{aligned}$$

Computing the partial derivative we find

$$\begin{aligned}\frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\log\left\{\sigma_{\hat{j}\hat{d}}^{s_{i\hat{d}}}(1-\sigma_{\hat{j}\hat{d}})^{1-s_{i\hat{d}}}\right\}\right] &= \frac{\frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\sigma_{\hat{j}\hat{d}}^{s_{i\hat{d}}}(1-\sigma_{\hat{j}\hat{d}})^{1-s_{i\hat{d}}}\right]}{\sigma_{\hat{j}\hat{d}}^{s_{i\hat{d}}}(1-\sigma_{\hat{j}\hat{d}})^{1-s_{i\hat{d}}}}\\ &= \frac{s_{i\hat{d}}\sigma_{\hat{j}\hat{d}}^{s_{i\hat{d}}-1}(1-\sigma_{\hat{j}\hat{d}})^{1-s_{i\hat{d}}} - (1-s_{i\hat{d}})(1-\sigma_{\hat{j}\hat{d}})^{-s_{i\hat{d}}}\sigma_{\hat{j}\hat{d}}^{s_{i\hat{d}}}}{\sigma_{\hat{j}\hat{d}}^{s_{i\hat{d}}}(1-\sigma_{\hat{j}\hat{d}})^{1-s_{i\hat{d}}}}\\ &= \frac{s_{i\hat{d}}}{\sigma_{\hat{j}\hat{d}}} - \frac{(1-s_{i\hat{d}})}{(1-\sigma_{\hat{j}\hat{d}})}\\ &= \frac{s_{i\hat{d}} - \sigma_{\hat{j}\hat{d}}}{\sigma_{\hat{j}\hat{d}}(1-\sigma_{\hat{j}\hat{d}})}\ .\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \frac{s_{i\hat{d}} - \sigma_{\hat{j}\hat{d}}}{\sigma_{\hat{j}\hat{d}}(1-\sigma_{\hat{j}\hat{d}})} &= 0 \\ \sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \left(s_{i\hat{d}} - \sigma_{\hat{j}\hat{d}}\right) &= 0\end{aligned}$$

which gives the $\sigma_{\hat{j}\hat{d}}$ estimate update

$$\sigma_{\hat{j}\hat{d}} = \frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] s_{i\hat{d}}}{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right]} \; .$$

A.1.3 Maximization for $\mu_{\hat{j}\hat{d}\hat{r}}$

Optimization with respect to a particular $\mu_{\hat{j}\hat{d}\hat{r}}$ is constrained by $\sum_{r=1}^{R} \mu_{\hat{j}\hat{d}r} = 1$. Using the Lagrange multiplier $\lambda_{\hat{j}\hat{d}}$, the optimality condition of Equation 3.1 with respect to this parameter is

$$\frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \log\left\{p(\mathbf{s}_i, \mathbf{m}_i \mid \boldsymbol{\Theta})\right\} + \lambda_{\hat{j}\hat{d}}\left(\sum_{r=1}^{R} \mu_{\hat{j}\hat{d}r} - 1\right)\right] = 0$$

which simplifies to

$$\begin{aligned}\frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \log\{p(\mathbf{m}_i \mid c_j)\} + \lambda_{\hat{j}\hat{d}} \sum_{r=1}^{R} \mu_{\hat{j}\hat{d}r}\right] &= 0 \\ \frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \log\left\{\prod_{d=1}^{D_M}\prod_{r=1}^{R} \mu_{jdr}^{[m_{id}=q_r]}\right\} + \lambda_{\hat{j}\hat{d}} \sum_{r=1}^{R} \mu_{\hat{j}\hat{d}r}\right] &= 0 \\ \frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \sum_{d=1}^{D_M}\sum_{r=1}^{R} \log\left\{\mu_{jdr}^{[m_{id}=q_r]}\right\} + \lambda_{\hat{j}\hat{d}} \sum_{r=1}^{R} \mu_{\hat{j}\hat{d}r}\right] &= 0 \\ \sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)}\right] \frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\log\left\{\mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}}==q_{\hat{r}}]}\right\}\right] + \lambda_{\hat{j}\hat{d}} &= 0 \; .\end{aligned}$$

Performing the computation of the partial derivative, we find

$$\frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\log \left\{ \mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]} \right\} \right] = \frac{\frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]} \right]}{\mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]}}$$

where we note that if $m_{i\hat{d}} = q_{\hat{r}}$ then

$$\frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]} \right] = \frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\mu_{\hat{j}\hat{d}\hat{r}} \right] = 1$$

and if $m_{i\hat{d}} \neq q_{\hat{r}}$ then

$$\frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]} \right] = \frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} [1] = 0 \ .$$

Thus,

$$\frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]} \right] = [m_{i\hat{d}} = q_{\hat{r}}]$$

and therefore

$$\frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\log \left\{ \mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]} \right\} \right] = \frac{[m_{i\hat{d}} = q_{\hat{r}}]}{\mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]}} \ .$$

Continuing the computation, we find

$$\begin{aligned} \sum_{i=1}^{n} E_Z \left[z_{i\hat{j}} \ \middle| \ \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right] \frac{\partial}{\partial \mu_{\hat{j}\hat{d}\hat{r}}} \left[\log \left\{ \mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]} \right\} \right] + \lambda_{\hat{j}\hat{d}} &= 0 \\ \sum_{i=1}^{n} E_Z \left[z_{i\hat{j}} \ \middle| \ \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right] \frac{[m_{i\hat{d}} = q_{\hat{r}}]}{\mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]}} + \lambda_{\hat{j}\hat{d}} &= 0 \\ \sum_{i=1}^{n} E_Z \left[z_{i\hat{j}} \ \middle| \ \mathbf{s}_i, \mathbf{m}_i, \boldsymbol{\Theta}^{(t)} \right] [m_{i\hat{d}} = q_{\hat{r}}] + \lambda_{\hat{j}\hat{d}} \mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}} = q_{\hat{r}}]} &= 0 \ . \end{aligned}$$

To find $\lambda_{\hat{j}\hat{d}}$, we sum over all r

$$\sum_{r=1}^{R}\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right][m_{i\hat{d}} = q_r] + \sum_{r=1}^{R} \lambda_{\hat{j}\hat{d}} \mu_{\hat{j}\hat{d}r}^{[m_{i\hat{d}}=q_r]} = 0$$

$$\sum_{r=1}^{R}\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right][m_{i\hat{d}} = q_r] + \lambda_{\hat{j}\hat{d}} = 0$$

which gives us

$$\lambda_{\hat{j}\hat{d}} = -\sum_{r=1}^{R}\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right][m_{i\hat{d}} = q_r]$$

So, we find the parameter estimate for $\mu_{\hat{j}\hat{d}\hat{r}}$ is

$$\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right][m_{i\hat{d}} = q_{\hat{r}}] \\ -\sum_{r=1}^{R}\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right][m_{i\hat{d}} = q_r]\mu_{\hat{j}\hat{d}\hat{r}}^{[m_{i\hat{d}}=q_{\hat{r}}]} = 0$$

and thus

$$\mu_{\hat{j}\hat{d}\hat{r}} = \frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{\sum_{r=1}^{R}\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right][m_{i\hat{d}} = q_r]} \;.$$

A.2 Single-tier Hidden-Parent Sublineage Model Updates

This section describes in detail the derivations required for the expectation-maximization algorithm as applied to the single-tier model with hidden parents, originally described in Section 3.1.2.

A.2.1 Maximization for $\alpha_{\hat{j}}$

The of the maximization of $\alpha_{\hat{j}}$ proceeds as in Section A.1.1, resulting in

$$\alpha_{\hat{j}} = \frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{n} \;.$$

A.2.2 Maximization for $\sigma_{\hat{j}\hat{d}}$

The optimality condition of Equation 3.1 with respect to the parameter $\sigma_{\hat{j}\hat{d}}$ is

$$\frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] \log\left\{p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right\}\right] = 0,$$

which simplifies to

$$\begin{aligned}\frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} E_Z\left[z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] \log\{p(\mathbf{s}_i \mid c_j)\}\right] &= 0 \\ \sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] \frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\log\left\{p(\mathbf{s}_i \mid c_j)\right\}\right] &= 0 \;.\end{aligned}$$

Computing the partial derivative we find

$$\begin{aligned}\frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\left[\log\left\{p(s_{i\hat{d}} \mid c_{\hat{j}})\right\}\right] &= \frac{\partial}{\partial \sigma_{\hat{j}\hat{d}}}\Bigg[\log\Bigg\{\Big(\eta_{11}\sigma_{\hat{j}\hat{d}} + \eta_{10}(1-\sigma_{\hat{j}\hat{d}})\Big)^{s_{i\hat{d}}} \\ &\qquad\qquad \Big(\eta_{01}\sigma_{\hat{j}\hat{d}} + \eta_{00}(1-\sigma_{\hat{j}\hat{d}})\Big)^{1-s_{i\hat{d}}}\Bigg\}\Bigg] \\ &= \frac{s_{i\hat{d}}\left(\eta_{11}-\eta_{10}\right)}{\eta_{11}\sigma_{\hat{j}\hat{d}} + \eta_{10}(1-\sigma_{\hat{j}\hat{d}})} + \frac{\left(1-s_{i\hat{d}}\right)\left(\eta_{01}-\eta_{00}\right)}{\eta_{01}\sigma_{\hat{j}\hat{d}} + \eta_{00}(1-\sigma_{\hat{j}\hat{d}})} \\ &= \frac{s_{i\hat{d}}\left(\eta_{11}-\eta_{10}\right)}{\eta_{11}\sigma_{\hat{j}\hat{d}} + \eta_{10}(1-\sigma_{\hat{j}\hat{d}})} + \frac{\left(1-s_{i\hat{d}}\right)\left(\eta_{11}-\eta_{10}\right)}{\eta_{10}-1+\sigma_{\hat{j}\hat{d}}(\eta_{11}-\eta_{10})} \\ &= \frac{\left(\eta_{11}-\eta_{10}\right)\Big(\eta_{10}-s_{i\hat{d}}+\sigma_{\hat{j}\hat{d}}(\eta_{11}-\eta_{10})\Big)}{(\eta_{11}\sigma_{\hat{j}\hat{d}} + \eta_{10}(1-\sigma_{\hat{j}\hat{d}}))(\eta_{10}-1+\sigma_{\hat{j}\hat{d}}(\eta_{11}-\eta_{10}))}\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] \frac{\left(\eta_{11}-\eta_{10}\right)\Big(\eta_{10}-s_{i\hat{d}}+\sigma_{\hat{j}\hat{d}}(\eta_{11}-\eta_{10})\Big)}{(\eta_{11}\sigma_{\hat{j}\hat{d}} + \eta_{10}(1-\sigma_{\hat{j}\hat{d}}))(\eta_{10}-1+\sigma_{\hat{j}\hat{d}}(\eta_{11}-\eta_{10}))} &= 0 \\ \sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]\left(\eta_{11}-\eta_{10}\right)\Big(\eta_{10}-s_{i\hat{d}}+\sigma_{\hat{j}\hat{d}}(\eta_{11}-\eta_{10})\Big) &= 0\end{aligned}$$

Solving for $\sigma_{\hat{j}\hat{d}}$, we compute

$$\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] \sigma_{\hat{j}\hat{d}}(\eta_{11} - \eta_{10})^2 =$$

$$\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] (\eta_{11} - \eta_{10})\,(s_{i\hat{d}} - \eta_{10})$$

and get the $\sigma_{\hat{j}\hat{d}}$ estimate update

$$\sigma_{\hat{j}\hat{d}} = \frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] (s_{i\hat{d}} - \eta_{10})}{(\eta_{11} - \eta_{10}) \sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}$$

A.2.3 Maximization for $\mu_{\hat{j}\hat{d}\hat{r}}$

The parameter update for $\mu_{\hat{j}\hat{d}\hat{r}}$ proceeds as in Section A.1.3, resulting in

$$\mu_{\hat{j}\hat{d}\hat{r}} = \frac{\sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{\sum_{r=1}^{R} \sum_{i=1}^{n} E_Z\left[z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] [m_{i\hat{d}} = q_r]} \; .$$

A.3 Multi-tier Hidden-Parent Model Updates

This section describes the derivations required for the EM algorithm as applied to the multi-tier model with hidden parents, originally described in Section 3.1.3.

A.3.1 Maximization for $\alpha_{\hat{u}}$

Optimization with respect to a particular $\alpha_{\hat{j}}$ is constrained by $\sum_{u=1}^{k_A} \alpha_u = 1$. Thus we have the optimality condition of Equation 3.3 with respect to the parameter $\alpha_{\hat{u}}$:

$$\frac{\partial}{\partial \alpha_{\hat{u}}} \left[\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W\left[x_{iu} y_{iv} z_{ij} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] \log\left\{p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})\right\} \right.$$

$$\left. + \lambda \left(\sum_{u=1}^{k_A} \alpha_u - 1 \right) \right] = 0 \; .$$

Breaking the probability $p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta})$ up into

$$p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) = p(a_u)p(b_v \mid a_u)p(c_j \mid b_v)p(\mathbf{s}_i \mid c_j)p(\mathbf{m}_i \mid c_j)$$

and recalling that $p(a_u) = \alpha_u$, we continue the computation of the optimality condition:

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{i\hat{u}} y_{iv} z_{ij} \ \middle|\ \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] \frac{\partial}{\partial \alpha_{\hat{u}}} \left[\log \{ \alpha_{\hat{u}} \} \right] + \lambda = 0 \\
\frac{\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{i\hat{u}} y_{iv} z_{ij} \ \middle|\ \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right]}{\alpha_{\hat{u}}} + \lambda = 0 \\
\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{i\hat{u}} y_{iv} z_{ij} \ \middle|\ \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] + \alpha_{\hat{u}} \lambda = 0
\end{aligned}$$

To find λ, we sum up over all $\partial / \partial \alpha_{\hat{u}}$:

$$\sum_{u=1}^{k_A} \sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{iu} y_{iv} z_{ij} \ \middle|\ \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] + \sum_{u=1}^{k_A} \alpha_u \lambda = 0$$

which gives the value of λ as

$$\begin{aligned}
\lambda &= -\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{iu} y_{iv} z_{ij} \ \middle|\ \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] \\
\lambda &= -n
\end{aligned}$$

where the calculations follow as we note that $\sum_{j=1}^{k} \alpha_j = 1$ (by constraint) and for particular i,

$$\sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{iu} y_{iv} z_{ij} \ \middle|\ \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] = \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B \\ j=1,\ldots,k_C}} p(a_u, b_v, c_j \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}) = 1 \ .$$

Thus,

$$\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{i\hat{u}} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] - \alpha_{\hat{u}} n = 0$$

and our final $\alpha_{\hat{u}}$ parameter maximization update is

$$\alpha_{\hat{u}} = \frac{\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{i\hat{u}} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right]}{n} .$$

A.3.2 Maximization for $\beta_{\hat{v}\hat{u}}$

Optimization with respect to a particular $\beta_{\hat{v}\hat{u}}$ is constrained by $\sum_{v=1}^{k_B} \beta_{v\hat{u}} = 1$ Thus we have the optimality condition of Equation 3.3 with respect to the parameter $\beta_{\hat{v}\hat{u}}$:

$$\frac{\partial}{\partial \beta_{\hat{v}\hat{u}}} \left[\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W \left[x_{iu} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] \log \left\{ p(\mathbf{s}_i, \mathbf{m}_i \mid \mathbf{\Theta}) \right\} \right.$$
$$\left. + \lambda_{\hat{u}} \left(\sum_{v=1}^{k_B} \beta_{v\hat{u}} - 1 \right) \right] = 0$$

or equivalently

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{k_C} E_W \left[x_{i\hat{u}} y_{i\hat{v}} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] \frac{\partial}{\partial \beta_{\hat{v}\hat{u}}} \left[\log \left\{ p(b_v \mid a_u) \right\} \right] + \lambda_{\hat{u}} &= 0 \\ \frac{\sum_{i=1}^{n} \sum_{j=1}^{k_C} E_W \left[x_{i\hat{u}} y_{i\hat{v}} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right]}{\beta_{\hat{v}\hat{u}}} + \lambda_{\hat{u}} &= 0 \\ \sum_{i=1}^{n} \sum_{j=1}^{k_C} E_W \left[x_{i\hat{u}} y_{i\hat{v}} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] + \beta_{\hat{v}\hat{u}} \lambda_{\hat{u}} &= 0 \end{aligned}$$

To find $\lambda_{\hat{u}}$, we sum up over all v:

$$\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W\left[x_{i\hat{u}} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] + \sum_{v=1}^{k_B} \beta_{v\hat{u}} \lambda_{\hat{u}} = 0$$

which results in the $\lambda_{\hat{u}}$

$$\lambda_{\hat{u}} = -\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W\left[x_{i\hat{u}} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]$$

where the calculations follow as we note that $\sum_{v}^{k_B} \beta_{vu} = 1$ (by constraint).

Thus,

$$\sum_{i=1}^{n} \sum_{j=1}^{k_C} E_W\left[x_{i\hat{u}} y_{i\hat{v}} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] \\ -\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W\left[x_{i\hat{u}} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right] \beta_{\hat{v}\hat{u}} = 0$$

and we have the resulting $\beta_{\hat{v}\hat{u}}$ parameter update

$$\beta_{\hat{v}\hat{u}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k_C} E_W\left[x_{i\hat{u}} y_{i\hat{v}} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{\sum_{i=1}^{n} \sum_{\substack{v=1,\ldots,k_B \\ j=1,\ldots,k_C}} E_W\left[x_{i\hat{u}} y_{iv} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]} \, .$$

A.3.3 Maximization for $\gamma_{\hat{j}\hat{v}}$

Maximization for a particular $\gamma_{\hat{j}\hat{v}}$ proceeds exactly as the $\beta_{\hat{v}\hat{u}}$ case, giving us

$$\gamma_{\hat{j}\hat{v}} = \frac{\sum_{i=1}^{n} \sum_{u=1}^{k_A} E_W\left[x_{iu} y_{i\hat{v}} z_{i\hat{j}} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]}{\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ j=1,\ldots,k_C}} E_W\left[x_{iu} y_{i\hat{v}} z_{ij} \;\middle|\; \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)}\right]} \, .$$

A.3.4 Maximization for $\sigma_{\hat{j}\hat{d}}$

Maximization for a particular $\sigma_{\hat{j}\hat{d}}$ occurs similarly to the way described in Section A.2.2, with the following result:

$$\sigma_{\hat{j}\hat{d}} = \frac{\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W \left[x_{iu} y_{iv} z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] (s_{i\hat{d}} - \eta_{10})}{(\eta_{11} - \eta_{10}) \sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W \left[x_{iu} y_{iv} z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right]} \; .$$

A.3.5 Maximization for $\mu_{\hat{j}\hat{d}\hat{r}}$

Maximization for a particular $\mu_{\hat{j}\hat{d}\hat{r}}$ occurs similarly to the way described in Section A.1.3, with the following result:

$$\mu_{\hat{j}\hat{d}\hat{r}} = \frac{\sum_{i=1}^{n} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} E_W \left[x_{iu} y_{iv} z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right]}{\sum_{r=1}^{R} \sum_{\substack{u=1,\ldots,k_A \\ v=1,\ldots,k_B}} \sum_{i=1}^{n} E_Z \left[x_{iu} y_{iv} z_{i\hat{j}} \mid \mathbf{s}_i, \mathbf{m}_i, \mathbf{\Theta}^{(t)} \right] [m_{i\hat{d}} = q_r]} \; .$$