Examining the sublineage structure of *Mycobacterium tuberculosis* complex strains with multiple-biomarker tensors

Cagri Ozcaglar[1], Amina Shabbeer[1], Scott Vandenberg[3], Bülent Yener[1], Kristin P. Bennett[1,2]
(1) Computer Science Department and (2) Mathematical Sciences Department, Rensselaer Polytechnic Institute
(3) Computer Science Department, Siena College
ozcagc2@cs.rpi.edu, shabba@cs.rpi.edu, vandenberg@siena.edu, yener@cs.rpi.edu, bennek@rpi.edu

***Abstract*—Strains of the *Mycobacterium tuberculosis* complex (MTBC) can be classified into coherent lineages of similar traits based on their genotype. We present a tensor clustering framework to group MTBC strains into sublineages of the known major lineages based on two biomarkers: spacer oligonucleotide type (spoligotype) and mycobacterial interspersed repetitive units (MIRU). We represent genotype information of MTBC strains in a high-dimensional array in order to include information about spoligotype, MIRU, and their coexistence using multiple-biomarker tensors. We use multiway models to transform this multidimensional data about the MTBC strains into two-dimensional arrays and use the resulting score vectors in a stable partitive clustering algorithm to classify MTBC strains into sublineages. We validate clusterings using cluster stability and accuracy measures, and find stabilities of each cluster. Based on validated clustering results, we present a sublineage structure of MTBC strains and compare it to the sublineage structures of SpolDB4 and MIRU-VNTR*plus*.**

***Index Terms*—Tuberculosis, *Mycobacterium tuberculosis* complex, multiway models, clustering, cluster validation**

I. Introduction

Tuberculosis (TB) is a bacterial disease caused by *Mycobacterium tuberculosis* complex (MTBC), and is a leading cause of death worldwide. In the United States, isolates from all TB patients are routinely genotyped by multiple biomarkers. The biomarkers include Spacer Oligonucleotide Types (spoligotypes), Mycobacterial Interspersed Repetitive Units - Variable Number Tandem Repeats (MIRU-VNTRs), IS6110 Restriction Fragment Length Polymorphisms (RFLP), Long Sequence Polymorphisms (LSPs) and Single Nucleotide Polymorphisms (SNPs).

Genotyping of MTBC is used to identify and distinguish MTBC into distinct lineages and/or sublineages that are useful for TB tracking and control and examining host-pathogen relationships [1]. The major lineages of MTBC are *M. africanum*, *M. canettii*, *M. microti*, *M. bovis*, *M. tuberculosis* subgroup Indo-Oceanic, *M. tuberculosis* subgroup Euro-American, *M. tuberculosis* subgroup East Asian (Beijing) and *M. tuberculosis* subgroup East-African Indian (CAS). These major lineages can be definitively characterized using LSPs [2], but typically only MIRU and spoligotypes are collected for the purpose of TB surveillance. Classification, similarity-search, and expert-rule based methods have been developed to correctly map isolates genotyped using MIRU and/or spoligotypes to the major lineages [3]–[5].

While sublineages of MTBC are routinely used in the TB literature, their exact definitions and names have not been clearly established. The SpolDB4 database contains 39,295 strains and their spoligotypes, with the vast majority of them labeled and classified into 62 sublineages [6], but many of these are considered to be "potentially phylogeographically-specific MTBC genotype families". Therefore, further analysis is needed to confirm these sublineages. The highly-curated MIRU-VNTR*plus* database, which focuses primarily on MIRU, defines 22 sublineages. New definitions of sublineages based on LSPs and SNPs are being discovered; e.g. the RD724 polymorphism corresponds to the previously defined SpolDB4 T2 sublineage, also known as the Uganda strain in MIRU-VNTR*plus* [7]. The SpolDB4 sublineages were created using only spoligotypes. Now large databases using both MIRU and spoligotypes exist. The United States Centers for Disease Control and Prevention (CDC) has gathered spoligotypes and MIRU isolates for over 37,000 patients. Well-defined TB sublineages based on MIRU and spoligotypes are critical for both TB control and research.

This study uses unsupervised multiway analysis to examine the sublineage structure of MTBC on the basis of spoligotype and MIRU patterns. The proposed method reveals structure not captured in SpolDB4 spoligotype families. When MIRU patterns are considered, SpolDB4 families that may be well supported by spoligotype signatures, become ambiguous, or may allow further subdivision. A key issue is how to combine spoligotype and MIRU into a single unsupervised learning model. A spoligotype-only tool, SPOTCLUST, was used to find MTBC sublineages using an unsupervised probabilistic model reflecting spoligotype evolution [8]. Existing phylogenetic methods can be readily applied to MIRU patterns, but specialized methods are needed to accurately capture how spoligotypes evolve. It is not known how to best combine spoligotype and MIRU to infer a phylogeny. The online tool www.miru-vntrplus.org determines lineages by using similarity search to a labeled database. The user must select the distance measure which is defined using spoligotypes and/or MIRU, possibly yielding different results.

In this study, we develop a tensor clustering framework for sublineage classification of MTBC strains labeled by major lineages. We generate multiple-biomarker tensors of MTBC strains and apply multiway models for dimensionality reduction. The model accurately captures spoligotype evolutionary dynamics by using contiguous deletions of spacers. The tensor transforms spoligotypes and MIRU into a new representation where traditional clustering methods apply (we use modified k-means clustering) without the users having to decide *a priori* how to combine spoligotype and MIRU patterns. Strains are clustered based on the transformed data without using any information from SpolDB4 families. Clustering results lead to the subdivision of major lineages of MTBC into groups with clear and distinguishable spoligotype and MIRU signatures. Comparison of the clusters with SpolDB4 families suggests dividing and merging some SpolDB4 families, while strongly validating others.

978-1-4244-8305-1/10/$26.00 ©2010 IEEE

II. Background

In this study, we used two genotyping methods, spoligotyping and MIRU-VNTR typing, to cluster MTBC strains. We generated high-dimensional arrays to represent genotype information of MTBC strains. We mapped these high-dimensional arrays to two-dimensional space using multiway models and used score matrices of these models as input to k-means clustering of MTBC strains. We validated the clustering results using cluster stability and accuracy measures. In this section, we give a brief background on genotyping, multiway modeling and clustering of MTBC strains.

A. Spoligotyping

Spoligotyping is a DNA fingerprinting method that exploits the polymorphisms in the direct repeat (DR) region of the MTBC genome to distinguish between strains. The DR region is a polymorphic locus in the genome of MTBC which comprises of direct repeats (36 bp), separated by unique spacer sequences of 36 to 41 bp [9]. The method uses 43 spacers, thus a spoligotype is typically represented by a 43-bit binary sequence. Zeros and ones in the sequence correspond to the absence and presence of spacers respectively. Mutations in the DR region involve deletion of contiguous spacers. To capture this evolution, we represent spoligotype deletions as a binary vector, where one indicates that a specific contiguous deletion occurs (i.e. a specified contiguous set of spacers are all absent) and zero means at least one spacer is present in that contiguous set of spacers.

B. MIRU-VNTR typing

MIRU is a homologous 46-100 bp DNA sequence dispersed within intergenic regions of MTBC, often as tandem repeats. Among the 41 identified mini-satellite regions on the MTBC genome, different subsets of size 12, 15 and 24 are proposed for standardization of MIRU genotyping [3]. In this study, we used 12-loci MIRU for genotyping MTBC. Thus, the MIRU pattern is represented as a vector of length 12, each entry representing the number of repeats in each MIRU locus.

C. Multiway analysis of biomarker tensor

The multiple-biomarker tensor captures three key properties of MTBC strains: spoligotype deletions, number of repeats in MIRU loci, and coexistence of spoligotype deletions with MIRU loci. This information is captured in a multidimensional array or tensor with three modes representing spoligotype deletions, MIRU patterns and strains. Mathematically, each strain is represented as the outer product of the binary spoligotype deletion vector and the MIRU pattern vector, which results in a biomarker kernel matrix. Kernel matrices of the same size for each strain form the multiple-biomarker tensor. Multiway models analyze tensors by decomposing multiway arrays into two-way arrays. In this study, we use two common multiway models, PARAFAC and Tucker3. Dimensionality reduction on the tensor data using multiway models returns a score vector for each MTBC strain, which is used to measure similarities and corresponding distances between strains in a clustering algorithm. This is a key property of the algorithm since we don't know *a priori* how to measure evolutionary distance between isolates genotyped by MIRU typing and spoligotyping.

III. Methods

Clustering MTBC strains using multiple biomarkers consists of a sequence of steps. First, we generate a tensor with one mode representing the strains to be clustered, and two other modes representing the two biomarkers. Second, we apply multiway models on the strain mode of the tensor to get a score matrix of strains. Third, we use this score matrix to decide similarity between strains, and cluster them using a stable version of k-means. In the final step, we evaluate the results of clusterings using cluster validity indices to select the best k. We outline the steps of the clustering framework in this section.

A. Datasets

The dataset comprises of 6848 distinct MTBC strains as determined by spoligotype and 12-loci MIRU, labeled with major lineages and SpolDB4 families. The strains are mainly from the CDC dataset - a database collected by the CDC from 2004-2008 labeled with the major lineages [4]. We also used the MIRU-VNTR*plus* dataset which is labeled with SpolDB4 lineages. The original SpolDB4 labeled dataset contains only spoligotypes. We found all occurrences of these spoligotypes in the CDC dataset. In this way we constructed a database with spoligotype and MIRU patterns, with major lineages as determined by CDC, and sublineages as given in SpolDB4. In total, the dataset has 571 East Asian (Beijing), 508 East-African Indian (CAS), 4580 Euro-American, 1023 Indo-Oceanic, 64 *M. africanum* and 102 *M. bovis* strains. We created 6 datasets from the CDC+MIRU-VNTR*plus* dataset, one for each major lineage, and divided them into sublineages.

Multiple-Biomarker Tensor: The dataset is arranged as a three-way array with strains in the first mode, spoligotype deletions in the second mode, and MIRU patterns in the third mode. Each entry $A(i, j, k)$ in the array corresponds to the number of repeats in MIRU pattern k of strain i with spoligotype deletion j. If spoligotype deletion j does not exist in strain i, then the tensor entry $A(i, j, .)$ is 0. Thus strain datasets are formed as *strain* $\times$ *spoligotype deletion* $\times$ *MIRU pattern* tensors. Generation of these multiple-biomarker tensors from the biomarker information of each strain is shown in Figure 1. We represent spoligotype deletions with $\vec{s}$, where $s_i \in \{0, 1\}$ and $i \in \{1, .., n\}$ where n is the number of informative spoligotype deletions found using feature selection. All possible deletions are classified by their frequency as nonexistent, uncommon and common deletions. Then, every uncommon deletion that is a supersequence of a common deletion is removed. The numbers of spoligotype deletions used for each major lineage are as follows: East Asian (Beijing) 5, East-African Indian (CAS) 18, Euro-American 109, Indo-Oceanic 28, *M. africanum* 22, and *M. bovis* 34. We represent 12-loci MIRU with $\vec{m}$, where $m_j \in \{0, .., 9, \geq 9\}$ and $j \in \{1, .., 12\}$. Details of the multiple-biomarker tensor generation can be found in [10].

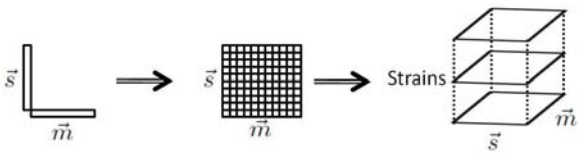

Fig. 1: Biomarker kernel matrix $\vec{s} \otimes \vec{m}$ for each strain forms multiple-biomarker tensor. $\vec{s}$ represents spoligotype deletions and $\vec{m}$ represents MIRU patterns.

B. Multiway modeling

We used the PARAFAC and Tucker3 techniques to model the three-way biomarker tensor. We determined the number of components for each model to ensure a bound on the explained variance of data.

1) Multiway models: We used PARAFAC and Tucker3 models to explain the tensor with high accuracy. Multiway modeling of multiple-biomarker tensors was carried out using the *n-way Toolbox* of MATLAB by Andersson et al. [11].

PARAFAC: PARAFAC is a generalization of SVD to multiway data [12]. A 3-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ is modeled by an R-component PARAFAC model as follows:

$$\underline{\mathbf{X}}_{ijk} = \sum_{r=1}^{R} \underline{\mathbf{G}}_{rrr} \mathbf{A}_{ir} \mathbf{B}_{jr} \mathbf{C}_{kr} + \underline{\mathbf{E}}_{ijk} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, $\mathbf{C} \in \mathbb{R}^{K \times R}$ are component matrices of first, second and third mode respectively. $\underline{\mathbf{G}} \in \mathbb{R}^{R \times R \times R}$ is the core array and $\underline{\mathbf{E}} \in \mathbb{R}^{I \times J \times K}$ is the residual term containing all unexplained variation.

Tucker3: Tucker3 is an extension of bilinear factor analysis to multiway datasets [13]. A 3-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ is modeled by a (P, Q, R)-component Tucker3 model as follows:

$$\underline{\mathbf{X}}_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} \underline{\mathbf{G}}_{pqr} \mathbf{A}_{ip} \mathbf{B}_{jq} \mathbf{C}_{kr} + \underline{\mathbf{E}}_{ijk} \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$, $\mathbf{C} \in \mathbb{R}^{K \times R}$ are component matrices of first, second and third mode. $\underline{\mathbf{G}} \in \mathbb{R}^{P \times Q \times R}$ is the core array and $\underline{\mathbf{E}} \in \mathbb{R}^{I \times J \times K}$ is the residual term.

2) Model validation: A multiway model is appropriate if adding more components to any mode does not improve the fit considerably. We used the core consistency diagnostic (CORCONDIA) to determine the number of components of the PARAFAC model [14]. The core consistency diagnostic measures the similarity of the core array $\underline{\mathbf{G}}$ of the model and the superdiagonal array of ones. As a rule of thumb, Bro et al. suggests that a core consistency above 90% implies an appropriate model [14].

In order to determine the number of components of the Tucker3 model, we started by fitting a Tucker3 model to the tensor with the same number of components. We picked the number of components that explains the variance of the data with close to 100% accuracy. Then we decreased the number of components until the most important factor combinations are found that explain over 90% of the variance of the data. The validated number of components along with core consistency values for PARAFAC models and explained variance for Tucker3 models are included in Table I.

Major Lineage	PARAFAC		Tucker3	
	# Components	Core Consistency	# Components	Variance
M. africanum	3	94.79	[4 4 3]	95.66
M. bovis	2	100.00	[7 6 4]	95.05
East Asian (Beijing)	2	100.00	[3 4 2]	93.09
East-African Indian (CAS)	2	100.00	[11 10 4]	97.23
Indo-Oceanic	4	94.32	[15 13 5]	95.55
Euro-American	14	99.03	[14 13 5]	89.77

TABLE I: Number of components used in PARAFAC and Tucker3 model to fit the tensors for the datasets to be clustered. We used core consistency diagnostic to validate PARAFAC models and percentage of explained variance to validate Tucker3 models.

C. Clustering algorithm

We developed the `kmeans_mtimes_seeded` algorithm, a modified version of the k-means algorithm, to group MTBC strains based on the score matrices of the multiway models. K-means is a commonly used clustering algorithm with two weaknesses: 1) Initial centroids are chosen randomly, 2) The objective value of k-means, measured as within-cluster sum of squares, may converge to local minima, rather than finding the global minimum. We solve these problems with two improvements: 1) Initial centroids are chosen by careful seeding, using a heuristic called `kmeans++`, suggested by Arthur et al. [15]. Let $D(x)$ represent the shortest Euclidean distance from data point x to the closest center already chosen. `kmeans++` chooses a new centroid at each step such that the new centroid is furthest from all chosen centroids. 2) The local minima problem is partially solved by repeating the k-means algorithm multiple times and getting the run with minimum objective value. The `kmeans_mtimes_seeded` algorithm is more stable than the k-means algorithm, and produces more accurate results. The number of clusters, k, is selected as described below. Details of `kmeans_mtimes_seeded` algorithm are included in [10].

D. Cluster Validation

Clustering results for the MTBC strains are evaluated to determine the best k and compare it with existing sublineages using cluster validity indices. We used best-match stability to pick the most stable clusterings. In case of a tie in average best-match stability, we used DD-weighted gap statistic or F-measure for cluster validation [16].

Best-Match Stability: The stability of a clustering is found by the distribution of pairwise similarities between clusterings of subsamples of the data. We use best-match stability suggested by Hopcroft et al. [17]. The algorithm clusters the data multiple times, and compares the reference cluster to the model clusterings. The stability of each cluster is calculated by finding the average best match between this cluster and the clusters identified using model clusterings. Given two sets C and $C^{'}$, the match value is:

$$match(C, C^{'}) = \frac{\mid C \cap C^{'} \mid}{\max(\mid C \mid, \mid C^{'} \mid)}$$

and we define the best-match stability of cluster C compared to a clustering of strains into k clusters as:

$$best\_match(C, \bigcup_{i=1}^{k} C_i) = \max_{i=1,..,k} match(C, C_i)$$

High average best-match values denote that the two clusters have many strains in common and are of roughly the same size.

DD-weighted Gap Statistic: We used the DD-weighted gap statistic developed by Yan et al. to validate clusterings and pick the correct number of clusters [18]. This is an improved version of the gap statistic suggested by Tibshirani et al., and it measures within-cluster homogeneity. We used uniform distribution over a box aligned with the principal components of the dataset as the reference distribution, referred to as DDgap/PC. This measure can also find the hierarchical structure of a dataset if it exists.

F-measure: The F-measure is the harmonic mean of the precision and recall of a clustering with respect to a reference

clustering. We use it to evaluate how similar the tensor sublineages are to the SpolDB4 families. According to the contingency table in Table II, precision, recall and F-measure are defined as follows:

$$P = \frac{a}{a+c} \qquad R = \frac{a}{a+b} \qquad F = \frac{2PR}{P+R}$$

Since the F-measure combines precision and recall of clustering results, it has proven to be a successful metric.

	Same cluster	Different clusters
Same class	a	b
Different classes	c	d

TABLE II: Contingency table. Given that there are n data points in the dataset, the following condition holds: $a+b+c+d = \binom{n}{2}$.

IV. Results

We subdivide each of the major lineages of MTBC into sublineages using multiple-biomarker tensors. For each major lineage, we generated the multiple-biomarker tensor using spoligotypes and MIRUs and applied multiway models to identify putative sublineages of each major lineage. To evaluate the resulting clusters, we compare them with the published SpolDB4 families for each major lineage dataset. The results are summarized in Table III. For each lineage, results show that the tensor approach finds highly stable sublineages (the best-match stability is $\geq$85%) and that the number of sublineages found using tensors is close but not always identical to the number of SpolDB4 families.

The F-measures range from 57% to 87% indicating that the sublineages found by the tensor only partially overlap with those of SpolDB4. Recall that the SpolDB4 families were created by expert analysis using only spoligotypes and that analysis by alternative biomarkers such as SNP and LSP has led to alternative definitions of MTBC sublineages. The tensor sublineages are based on spoligotype and MIRU, thus in some cases the tensor divides SpolDB4 families due to difference in MIRU even if the spoligotypes match. In other cases, the tensor analysis merges together the SpolDB4 families because the collective spoligotypes and MIRU are very close. In some cases, the tensor analysis almost exactly reproduces a SpolDB4 family providing strong support for the existence of these families with no expert guidance. Thus multiway analysis of MTBC strains of each major lineage with multiple biomarkers leads to new sublineages and reaffirms existing ones. Further insight can be obtained by examining the putative sublineages for each major lineage. The confusion matrix, PCA plot, spoligotype signatures and MIRU signatures for all lineages can be found in the full length technical report [10].

A. Sublineage structure of M. africanum

The tensor methodology used Tucker3 to construct four distinct sublineages for *M. africanum*. Figure 2 shows heat maps representing the spoligotype and MIRU signatures for each of the tensor sublineages with white indicating 0 probability and black indicating probability of 1. Table IV gives the stability of each sublineage and the correspondence between the tensor sublineages and the SpolDB4 families. The four sublineages are quite distinct as shown by the stability of 1 for each sublineage and the clear separation of the four sublineages in the PCA plot in Figure 2.

The tensor sublineages strongly support the existence of the SpolDB4 AFRI_1, AFRI_2 and AFRI_3 families and show that the AFRI family is composed of these three families. With an

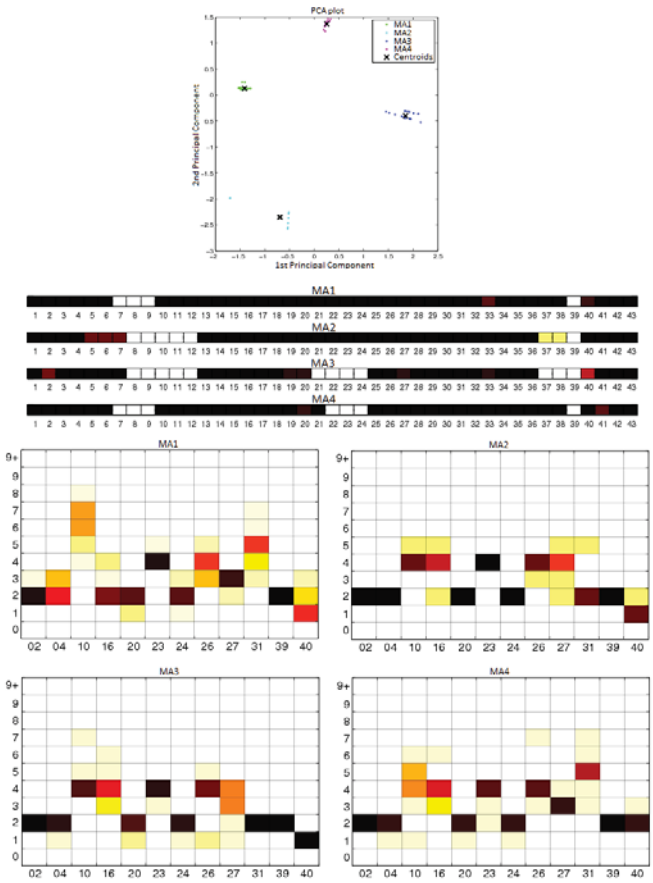

Fig. 2: PCA plot of clustering, spoligotype signatures and MIRU signatures of tensor sublineages of *M. africanum* strain dataset.

	MA1	MA2	MA3	MA4
Stability	1	1	1	1
AFRI	2	1	5	0
AFRI_1	21	0	0	16
AFRI_2	0	0	12	0
AFRI_3	0	6	1	0

TABLE IV: Confusion matrix for 64 distinct *M. africanum* strains showing the correspondence between the SpolDB4 families and tensor sublineages. The stability of each of the tensor sublineages is given in the second row.

F-measure of 66%, the tensor sublineages differ markedly from the SpolDB4 families for the *M. africanum* lineage. The AFRI family results largely explain this difference – AFRI is spread across three tensor sublineages. Disregarding AFRI, sublineages MA2 and MA3 match families AFRI_3 and AFRI_2 respectively. Interestingly, AFRI_1 is further subdivided into sublineages MA1 and MA4. The spoligotypes in MA1 and MA4 differ by only one contiguous deletion of spacers 22 through 24, but their MIRU signature clearly distinguishes them especially in MIRU loci 10, 12 and 40. The tensor indicates that the AFRI sublineage classification defines somewhat generic *M. africanum* strains that can be distinctly placed in the groups MA1 (part of AFRI_1), MA4 (other part of AFRI_1), MA2 (AFRI_3) and MA3 (AFRI_2).

The MIRU-VNTR*plus* labels, determined on the basis of LSPs

Major Lineage	# SpolDB4 families	# Tensor sublineages	F-measure	Average best-match stability
M. africanum	4	4	0.66	1
M. bovis	5	3	0.71	1
East Asian (Beijing)	2	5	0.87	1
East-African Indian (CAS)	4	3	0.82	1
Indo-Oceanic	13	11	0.57	0.90
Euro-American	33	33	0.61	0.85

TABLE III: Number of SpolDB4 families and number of tensor clusters for each major lineage. F-measure and best-match stability values assess the agreement of the sublineages to the SpolDB4 families and the certainty of tensor sublineages respectively.

	MA1	MA2	MA3	MA4
West African 1	0	0	5	0
West African 2	21	0	0	16
Unspecified	2	7	13	0

TABLE V: Confusion matrix for 64 distinct *M. africanum* strains showing the correspondence between the West African 1 and 2 sublineages and tensor sublineages. For data not from MIRU-VNTRplus, the lineage is indicated as unspecified.

indicate that there are two sublineages, West African 1 and West African 2 within *M. africanum*. Table V indicates the correspondence between the tensor sublineages and MIRU-VNTR*plus* labels. MA1 and MA4 clearly correspond to West African 2 and MA3 corresponds to West African 1. There are no data labeled by MIRU-VNTR*plus* in MA2, but we speculate that it is West African 1 since MA2 and MA3 have more closely related MIRU and spoligotype signatures than MA2 and MA1.

B. Sublineage structure of M. bovis

The tensor methodology used PARAFAC to construct 3 sublineages for *M. bovis*, MB1, MB2 and MB3, while the dataset contains 5 SpolDB4 families, BOV, BOVIS1, BOVIS1_BCG, BOVIS2 and BOVIS3. All the clusters have perfect stability and are well distinguished in the PCA plot [10]. Much like the *M. africanum* SpolDB4 AFRI family, the BOV family defines a generic *M. bovis* that spreads across all three tensor sublineages. Disregarding BOV, MB1 consists of all of BOVIS1 and BOVIS1_BCG. Since BOVIS1_BCG is the attenuated bacillus Calmette-Guérin (BCG) vaccine strain, it is difficult to distinguish it from BOVIS1 using only MIRU and spoligotypes. Therefore, the merger of BOVIS1 and BOVIS1_BCG makes genetic sense. Disregarding BOV, the MB2 and MB3 sublineages exactly match the SpolDB4 families BOVIS3 and BOVIS2 respectively.

C. Sublineage structure of East Asian (Beijing)

The tensor methodology used PARAFAC to construct five distinct sublineages for East Asian denoted B1 through B5. The variability in the spoligotypes of East Asian is limited to spacers 35 through 43 since all East Asian strains have spacers 1 to 34 absent. Since the SpolDB4 classification is based only on spoligotypes, the limited variability allows only two families, BEIJING and BEIJING-LIKE. The tensor cleanly subdivides BEIJING into three sublineages B1, B4 and B5 all with stability 1. Spoligotype signatures of these sublineages differ, and MIRU signature of sublineage B5 is clearly distinct in MIRU locus 40. The tensor subdivides the BEIJING-LIKE into sublineages B2 and B3 each with distinct spoligotype signatures. Thus the tensor strongly supports the existence of BEIJING and BEIJING-LIKE families, but also suggests that they can be further subdivided.

D. Sublineage structure of East-African Indian (CAS)

The tensor methodology used PARAFAC to construct three distinct sublineages for East-African Indian (also known as CAS) denoted C1, C2 and C3, while the dataset has four SpolDB4 lineages CAS, CAS1_DELHI, CAS1_KILI and CAS2. All sublineages are highly stable with stability 1. Much like with AFRI and BOV, the generic CAS family was divided across C1, C2, and C3 sublineages. Disregarding CAS, C1 only contains CAS1_DELHI and C3 only contains CAS2. C2 contains all of CAS1_KILI. C2 also contains 6 CAS1_DELHI strains, but the vast majority (327 strains) of CAS1_DELHI fall in C1. Variabilities in MIRU loci 10, 26, and 40 are key to defining differences in the sublineages along with distinct deletion patterns in the spoligotypes.

E. Sublineage structure of Indo-Oceanic

The tensor methodology used PARAFAC to construct eleven distinct sublineages for Indo-Oceanic denoted IO1 to IO11 while the dataset has thirteen SpolDB4 lineages. The EAI5 family acts much like the CAS, BOV and AFRI families, spreading across all the Indo-Oceanic sublineages except IO2 and IO5. The small MANU1 family also spreads across four sublineages. The existence of the MANU1 family has not been well established by other biomarkers. Disregarding these two troubling families, the tensor sublineages correspond closely to the SpolDB4 families. Specifically, the mapping between the most stable clusters (with sublineage stability) and the families are IO1 (.99) equals EAI3_IND, IO2 (1) equals ZERO, IO3 (.99) equals EAI2_NTB, IO4 (.98) equals a subset of EAI5, IO9 (.97) equals some EAI5 plus all of EAI8_MDG and some of EAI1_SOM, IO11 (.94) contains the vast majority of EAI1_SOM and EAI6_BDG1, and some of EAI5, and IO7 (.79) equals EAI4_VNM and EAI. EAI2_MANILLA is subdivided into three sublineages: IO8 (1) consisting of 241 strains, IO5 (.81) with 24 strains, and IO10 (.69) with 11 strains. While the spoligotype and MIRU signatures show that there are distinct EAI5 subgroups, the definition of the EAI5 and MANU1 groups are not well supported by the tensor analysis. They may represent a more general sublineage that is further subdivided. Distinct patterns are observable in the spoligotype and MIRU signatures for most of the lineages.

F. Sublineage structure of Euro-American

We used Tucker3 to find 33 sublineages for Euro-American denoted E1 to E33, the same number as the dataset which has 33 SpolDB4 lineages. Strains belonging to families H2, H37Rv, H4, LAM12_MAD1, T1 (Tuscany variant), T1_RUS2, T4, T5_MAD2 and T5_RUS1 are clustered in tensor sublineages E15, E24, E12, E8, E18, E6, E29, E29 and E18 respectively. In contrast, the

T1 family, an ancestral strain family, is distributed across 25 tensor sublineages, with most of the T1 strains in E29. Sublineage stability is above .90 for 18 tensor sublineages. Spoligotype and MIRU signatures of sublineages suggest either subdivision or merging of SpolDB4 families. For instance, tensor sublineages E2, E14 and E25 include T1 strains only. In addition to common spacer deletions of Euro-American strains, E2 lacks spacers 15 through 26, E14 lacks spacers 9 through 23 and E25 lacks spacers 3 through 12 and 14 through 18. This sublineage classification further subdivides the poorly-defined ancestor T1 family. Strains of LAM families on the other hand are grouped together in tensor sublineages E1 and E21. Prior studies have found that LAM Rio strains identified by SNPs are found in multiple SpolDB4 lineages [19]. Therefore, it is not surprising that use of the multiple biomarkers leads to subdivision or merging of some SpolDB4 families.

V. Conclusion

We developed a clustering framework which groups MTBC strains based on their spoligotype and MIRU information via multiple-biomarker tensors. We generated multiple-biomarker tensors for representation of high-dimensional biomarker information and used multiway models for dimensionality reduction. The multiway representation determines a transformation of the data that captures the similarities and differences between strains based on two distinct biomarkers. We clustered MTBC strains based on transformed data using improved k-means clustering and validated clustering results. We evaluated the sublineage structure of major lineages of MTBC and found similarities and clear distinctions in our subdivision of major lineages compared to the SpolDB4 classification. Simultaneous analysis of spoligotype and MIRU through multiple-biomarker tensors and clustering of MTBC strains lead to coherent sublineages of major lineages with clear and distinctive spoligotype and MIRU signatures.

The clustering framework used in this study can be further extended to find subgroups of MTBC strains based on other biomarkers such as RFLP and SNPs. We can use spoligotype and MIRU to group MTBC strains and compare them to labels derived from SNPs. Representation of MTBC genotype via multiple-biomarker tensors can also be extended to include 15-loci and 24-loci MIRU patterns. Moreover, more biomarkers can be used in the MTBC strain genotype representation. We can extend multiple-biomarker tensors and add a new mode for each biomarker added to the genotype representation of strains, e.g. RFLP. This would be a major advancement because there is no way to define a similarity measure between RFLPs of strains other than determining whether or not the patterns match exactly. Addition of new biomarkers will increase the number of modes of the multiple-biomarker tensor, while the multiway analysis methods remain the same.

Future work will involve using various biomarkers to group MTBC strains. Multiple-biomarker tensors with spoligotype, MIRU patterns, and RFLP in modes may lead to a clustering of MTBC strains which is comparable with lineages identified on the basis of SNPs. This flexible representation should enable identification of subgroups of MTBC strains based on nucleotide sequences in one of the modes. Since many subfamilies are clearly known and more biomarkers are being developed, the multiple-biomarker tensor can be used in supervised and even semi-supervised classification to build reliable classifiers of MTBC sublineages and can be used to enhance TB control, epidemiology and research.

Acknowledgment

This work was made possible by Dr. Lauren Cowan and Dr. Jeff Driscoll of the Centers for Disease Control and Prevention. This work was supported by NIH R01LM009731.

References

[1] S. Gagneux *et al.*, "Variable host-pathogen compatibility in *Mycobacterium tuberculosis*," *PNAS*, vol. 103, no. 8, pp. 2869–2873, 2006.

[2] S. Gagneux and P. M. Small, "Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development," *The Lancet Infectious Diseases*, vol. 7, no. 5, pp. 328 – 337, 2007.

[3] P. Supply *et al.*, "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. 4498–4510, 2006.

[4] 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.

[5] S. Ferdinand *et al.*, "Data mining of *Mycobacterium tuberculosis* complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families," *Research in Microbiology*, vol. 155, no. 8, pp. 647–654, 2004.

[6] K. Brudey *et al.*, "*Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology," *BMC Microbiology*, vol. 6, p. 23, 2006.

[7] B. Asiimwe, "Molecular characterization of *Mycobacterium tuberculosis* complex in Kampala, Uganda," Ph.D. dissertation, Makerere University, 2008.

[8] I. Vitol, J. Driscoll, B. Kreiswirth, N. Kurepina, and K. P. Bennett, "Identifying *Mycobacterium tuberculosis* complex strain families using spoligotypes," *Infection, Genetics and Evolution*, vol. 6, no. 6, pp. 491 – 504, 2006.

[9] J. Kamerbeek *et al.*, "Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology," *Journal of Clinical Microbiology*, vol. 35, no. 4, pp. 907–914, 1997.

[10] C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, and K. P. Bennett, "A clustering framework for *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors," Department of Computer Science, Rensselaer Polytechnic Institute, Tech. Rep. 10-08, 2010, available at http://www.cs.rpi.edu/research/pdf/10-08.pdf.

[11] C. A. Andersson and R. Bro, "The N-way toolbox for MATLAB," *Chemometrics and Intelligent Laboratory Systems*, vol. 52, no. 1, pp. 1 – 4, 2000.

[12] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 1, pp. 6–20, 2009.

[13] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[14] R. Bro and H. Kiers, "A new efficient method for determining the number of components in PARAFAC models," *Journal of Chemometrics*, vol. 17, no. 5, pp. 274–286, 2003.

[15] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[16] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Cluster validity methods: part I," *SIGMOD Rec.*, vol. 31, no. 2, pp. 40–45, 2002.

[17] J. Hopcroft *et al.*, "Natural communities in large linked networks," in *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 541–546.

[18] M. Yan and K. Ye, "Determining the number of clusters using the weighted gap statistic," *Biometrics*, vol. 63, no. 4, pp. 1031–7, 2007.

[19] A. L. Gibson *et al.*, "Application of sensitive and specific molecular methods to uncover global dissemination of the major RD^{Rio} sublineage of the Latin American-Mediterranean *Mycobacterium tuberculosis* spoligotype family," *J. Clin. Microbiol.*, vol. 46, no. 4, pp. 1259–1267, 2008.