Multiple-biomarker tensor analysis for tuberculosis lineage identification

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

1 Introduction

Tuberculosis (TB) is a bacterial disease caused by *Mycobacterium tuberculosis* complex (MTBC), and is a leading cause of death worldwide. 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). 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 are classified into 62 sublineages [2], but many of these are considered to be "potentially phylogeographically-specific MTBC genotype families". Therefore, further analysis is needed to confirm these sublineages.

In this study, we develop a tensor clustering framework for sublineage classification of MTBC strains labeled by major lineages based on multiple biomarkers, spoligotype and MIRU, which are the biomarkers typically collected for the purpose of TB surveillance. 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.

2 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. This stepwise clustering framework is outlined in Figure 1.

2.1 Multiple-biomarker tensor

The strain 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. 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 2. 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. We represent 12-loci MIRU with $\vec{m}$, where $m_j \in \{1,..,9,\geq 9\}$ and $j \in \{1,..,12\}$. Given multiple-biomarker tensor $\underline{\mathbf{X}}$, the entries can take the following values according to the genotype of the strain:

$$\underline{\mathbf{X}}_{ijk} = r = \begin{cases} 0 & \text{if strain } i \text{ lacks spoligotype deletion } j, \\ > 0 & \text{if strain } i \text{ has spoligotype deletion } j \text{ and MIRU loci } k \text{ has } r \text{ repeats.} \end{cases}$$

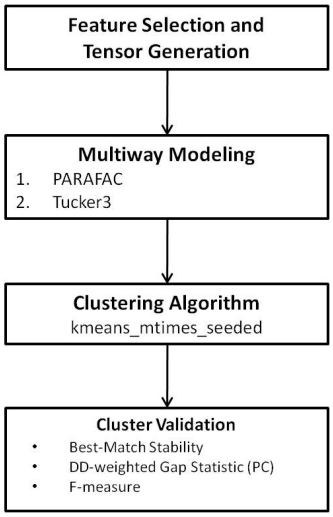

Fig. 1: Clustering framework of MTBC strains.

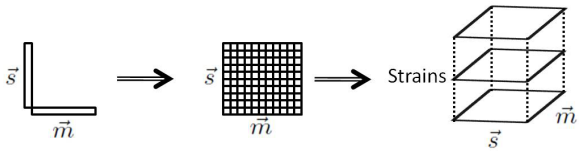

Fig. 2: Biomarker kernel matrix $\vec{s} \otimes \vec{m}$ for each strain forms multiple-biomarker tensor. $\vec{s}$ represents spoligotype deletions and $\vec{m}$ represents MIRU patterns.

Dataset: 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 [3]. 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.

2.2 Multiway modeling

We used 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.

2.2.1 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. [4].

PARAFAC: PARAFAC is a generalization of SVD to multiway data [5]. 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 [6]. 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.

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 1: 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.

2.2.2 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. As a rule of thumb, Bro et al. suggest that a core consistency above 90% implies an appropriate model [7]. 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. Validated number of components along with core consistency values for PARAFAC models and explained variance for Tucker3 models are included in Table 1.

2.3 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. We solved weaknesses of k-means with two improvements: 1) Initial centroids are chosen by careful seeding, using a heuristic called `kmeans++`, suggested by Arthur et al. [8]. 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. Details of `kmeans_mtimes_seeded` algorithm are included in [9].

2.4 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 for stability analysis of the clustering algorithm [10]. We used DD-weighted gap statistic and F-measure for cluster validation [11].

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 2: Number of SpolDB4 families and number of tensor sublineages 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.

3 Results

We subdivide each of the major lineages of MTBC into sublineages using multiple-biomarker tensors. Overall results for six major lineages are shown in Table 2. The F-measure values range from 57% to 87% indicating that the sublineages found by the tensor only partially overlap with those of SpolDB4. The four sublineages of *M. africanum*

strains found by tensor sublineages are quite distinct as shown the clear separation of the four sublineages in the PCA plot and biomarker signature in Figure 3. The tensor sublineages for all major lineages can be found in full length technical report [9].

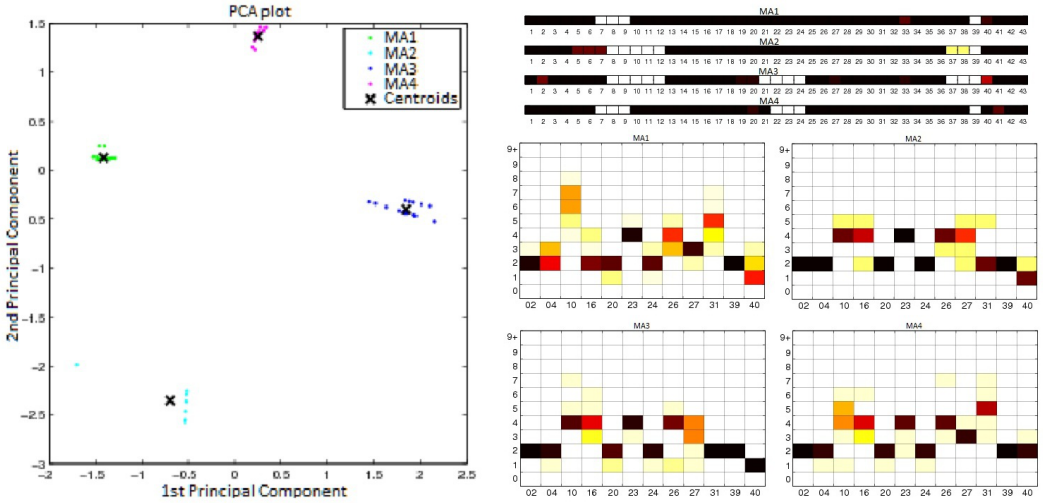

Fig. 3: PCA plot of clustering, spoligotype signatures and MIRU signatures of tensor sublineages of *M. africanum* strain dataset.

4 Conclusion

We developed a clustering framework which groups MTBC strains based on their spoligotype and MIRU information via multiple-biomarker tensors. 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 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. 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. Since many subfamilies are clearly known and more biomarkers are being developed, the multiple-biomarker tensor can be used in supervised classification to build reliable classifiers of MTBC sublineages and can be used to enhance TB control, epidemiology and research.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, and K. P. Bennett, "Examining the sublineage structure of *Mycobacterium tuberculosis* complex strains with multiple-biomarker tensors," Department of Computer Science, Rensselaer Polytechnic Institute, Tech. Rep., 2010. [Online]. Available: www.cs.rpi.edu/~ozcagc2/MultipleBiomarkerTensorsTechnicalReport.pdf

[10] 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.

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