Host-pathogen association analysis of tuberculosis patients via Unified Biclustering Framework

Cagri Ozcaglar*[1], Bülent Yener†[1], and Kristin P. Bennett‡[1,2]

[1]Computer Science Department, Rensselaer Polytechnic Institute
[2]Mathematical Sciences Department, Rensselaer Polytechnic Institute

1 Introduction

Tuberculosis (TB) is an airborne disease which is a leading cause of death worldwide. According to World Health Organization, one third of the human population is infected either latently or actively with TB [1]. *Mycobacterium tuberculosis* complex (MTBC) is the set of species which causes TB. MTBC isolates from TB patients are genotyped using multiple biomarkers for tracking TB transmission, TB control, and examining host-pathogen relationships.

Earlier studies have found associations between TB patients and the MTBC strains which infected them. Hirsh et al. showed that a TB patient's place of birth can be used to predict the geographic origin of the MTBC isolate [2]. Gagneux et al. defined the population structure of MTBC strains using six phylogeographic lineages and showed that these lineages are adapted to particular human populations defined by place of birth or risk factor [3]. Visual inspection via host-pathogen maps enable making inferences from patient data and strain lineages [4]. Although names of phylogeographic lineages imply an association between MTBC isolates and patients' place of birth, none of these studies combine genetic proximity between MTBC strains and spatial proximity between TB patients together. In this study, in addition to the distribution of MTBC isolates to their host's country of birth, we add genetic proximity, spatial proximity and time into domain knowledge of host-pathogen association analysis.

Multiple sources of information can be incorporated into data analysis via data fusion [5]. Recently, there has been considerable work on genomic data fusion [6–8]. In the TB context, Ozcaglar et al. built the tensor clustering framework (TCF) to cluster MTBC strains using multiple biomarkers simultaneously through genomic data fusion [9]. Genomic and phenomic data sources are also combined in earlier studies [10] via genome-phenome data fusion. However, there is no significant work on genome-phenome interactions of MTBC isolates and TB patients.

In this study, we present host-pathogen associations of tuberculosis by incorporating genetic proximity between MTBC strains, spatial proximity between TB patients, and time into domain knowledge via Unified Biclustering Framework (UBF). We simultaneously factorize multiple sources of information in various forms and obtain biclusters which represent host-pathogen pairs, while keeping pathogens genetically close in order to estimate most likely mutation events, and keeping hosts spatially close in order to estimate most likely transmission events. Based on factor matrices of hosts and pathogens, we generate the feature pattern similarity matrix of host-pathogen pairs, and find density-invariant biclusters. Finally, we select statistically

*ozcagc2@cs.rpi.edu
†yener@cs.rpi.edu
‡bennek@rpi.edu

significant biclusters among them and find the most stable host-pathogen associations. We also find host-pathogen associations within each major lineage. We evaluate biological relevance of statistically significant biclusters, confirm known host-pathogen associations, and propose new ones.

2 Background

In order to find relationships between MTBC isolates and TB patients, we uniquely identified them by their characteristics. We represented MTBC strains with a commonly used biomarker, spoligotype, and represented each patient with their country of birth. Finally, we stated the host-pathogen association analysis as a biclustering problem. Next, we give a brief background on spoligotyping, biclustering, and explain host-pathogen association analysis as a biclustering problem.

2.1 Spoligotyping

Spoligotyping is a DNA fingerprinting method of MTBC which exploits the polymorphism in the DR region consisting of 36 bp of direct repeats separated by 36 to 41 bp of spacers [11]. A spoligotype consists of 43 spacers, and it is represented as a 43-bit binary vector, where zeros represent absence of spacers and ones represent presence of spacers. Mutations in the DR region can result in loss of spacers, but not gain. This rule of irreversible mutation of spoligotypes is also known as contiguous deletion assumption [12,13].

2.2 Biclustering

Biclustering is a class of clustering algorithms which perform simultaneous clustering of rows and columns of a matrix. The term was first coined by Cheng and Church for gene expression data analysis [14]. Following them, many biclustering algorithms motivated mostly by bioinformatics applications are developed. These biclustering algorithms include spectral biclustering algorithm by Dhillon et al. [15] and Kluger et al. [16], Statistical-Algorithmic Method for Bicluster Analysis (SAMBA) by Tanay et al. [17], Coupled Two-Way Clustering (CTWC) by Getz et al. [18], Binary Inclusion-Maximal biclustering algorithm (BiMax) by Prelic et al. [19], and densely-connected biclustering (DECOB) by Colak et al. [20]. A great survey by Madeira et al. details biclustering and existing biclustering algorithms for biological data analysis [21].

2.3 Host-pathogen association analysis: a biclustering problem

Biclustering was initially motivated by gene expression data analysis in order to group genes into subsets of genes which are coexpressed under certain subsets of conditions. This is equivalent to finding submatrices in a gene expression matrix such that the submatrix entries follows a cohesive pattern under investigation. In the TB context, the genes of microarray data maps to spoligotypes of MTBC strains, and the conditions of microarray data maps to country of birth of TB patients. The resulting host-pathogen matrix of tuberculosis expresses the association level of a spoligotype to a country.

In the case where the original host-pathogen matrix is extended or concatenated with other matrices via data fusion, we use feature patterns for spoligotypes and countries. We first extract feature patterns for each spoligotype of MTBC strains and for each country of birth for TB patients. The association level of a spoligotype and a country is calculated as the cosine similarity of their feature pattern vectors. This final form of host-pathogen matrix of tuberculosis expresses association level of host-pathogen pairs, and is in the correct form to be analyzed via biclustering.

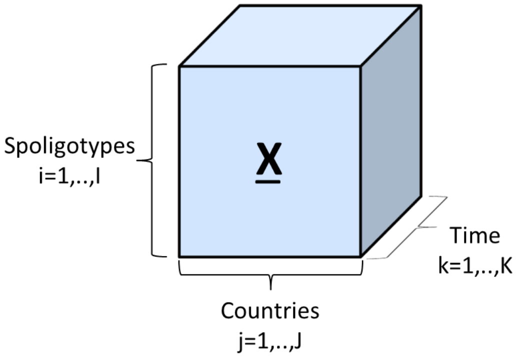


Figure 1: Host-pathogen tensor (HPT). The first mode represents spoligotypes, the second mode represents countries, and the third mode represents time. This HPT is of the form *Spoligotypes* × *Countries* × *Time.*

In the next section, we present the methods used for host-pathogen association analysis. We first give details about the patient dataset. Then, we present the calculation of genetic proximity matrix and spatial proximity matrix used in data fusion. Finally, we present the steps of Unified Biclustering Framework (UBF).

3 Methods

3.1 The dataset

The NYC dataset consists of 4876 patients in the United States diagnosed between 2001 and 2007. The spoligotype of MTBC strains and their host's country of birth are available in the dataset, along with the date of diagnosis. There are 858 unique spoligotypes in the original dataset. MTBC strains are labeled by major lineages based on their spoligotypes using Conformal Bayesian Network (CBN) model [22], and by KBBN sublineages using the Knowledge-based Bayesian Network (KBBN) model [23]. We refer to spoligotypes using shared type numbers, or SIT numbers using SITVITWEB database [24]. If the spoligotype is not assigned to an ST number by SITVITWEB, then we assign a unique UST number, where U denotes unknown ST. We first filter this data such that there are at least 2 patients from each country, and at least two patients infected with each strain. After filtering the dataset, there remains 4301 patients, 311 spoligotypes, and 104 countries. Using this filtered dataset, we construct the host-pathogen tensor (HPT) of the form *Spoligotypes* × *Countries* × *Time*. The final HPT is denoted as $\underline{\mathbf{X}} \in \mathbb{R}^{(I=311)\times(J=104)\times(K=7)}$. The host-pathogen tensor (HPT) is shown in Figure 1.

3.2 Distance matrices

In the host-pathogen tensor, the first mode represents pathogen attributes, in this case spoligotypes. Genetic proximity of spoligotypes can be found using genetic distance measures. Hosts with genetically close spoligotypes are more likely to be involved in the same mutation event. Similarly, the second mode represents host attributes, in this case country of birth. Proximity of countries can be found based on neighbourhood. Patients from close countries based on the proximity values are more likely to be involved in the same transmission event.

3.2.1 Genetic proximity matrix

Given 311 distinct spoligotypes, we define a genetic proximity measure between them. Mutation of spoligotypes is based on the Contiguous Deletion Assumption (CDA), which states that one or more contiguous spacers can be deleted in a mutation event, but not gained. Let s_i represent spoligotype i, and let $s_i \rightarrow s_j$ represent the mutation of spoligotype s_i into spoligotype s_j. Then, we define the CDA matrix, which summarizes contiguous deletion assumption, as follows:

$$\text{CDA}(s_i, s_j) = \begin{cases} \text{true}, & \text{if } s_i \rightarrow s_j \text{ or } s_j \rightarrow s_i \\ \text{false}, & \text{otherwise.} \end{cases}$$

Let $H(s_i, s_j)$ be the Hamming distance between spoligotypes s_i and s_j, as defined in [12]:

$$H\left(s_i, s_j\right) = \sum_{r=1}^{43} \mid s_{ir} - s_{jr} \mid$$

where s_{ir} represents the value of $r-th$ spacer of spoligotype s_i. Then, we define the genetic proximity matrix P_G as follows:

$$P_G(s_i, s_j) = \begin{cases} \dfrac{1}{1 + H(s_i, s_j)}, & \text{if } i \neq j,\ \text{CDA}(s_i, s_j),\ H(s_i, s_j) \leq 10 \\ 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

Genetic proximity matrix P_G has values inversely proportional to the Hamming distance between two spoligotypes, as long as the Hamming distance between them is at most 10. For spoligotype pairs with $H(s_i, s_j) > 10$, the genetic proximity is set to zero. As a result, genetic proximity matrix reflects the likelihood of two different pathogens being involved in the same mutation event.

3.2.2 Spatial proximity matrix

Given 104 countries, we first define the Country Neighbourhood Matrix (CNM). Given two countries C_i and C_j, the CNM is defined as follows:

$$\text{CNM}(C_i, C_j) = \begin{cases} 1, & \text{if } C_i \text{ and } C_j \text{ are neighbours} \\ 0, & \text{otherwise.} \end{cases}$$

Let $L(C_i, C_j)$ be the length of shortest path from C_i to C_j based on Dijkstra's shortest path algorithm on CNM [25]. Then, we define the spatial proximity matrix P_S as follows:

$$P_S(C_i, C_j) = \begin{cases} \dfrac{1}{1 + L(C_i, C_j)}, & \text{if } i \neq j,\ L(C_i, C_j) \leq 3 \\ 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

Spatial proximity matrix P_S has values inversely proportional to the length of shortest path between two countries, as long as the shortest-path length is at most 3. For country pairs with shortest-path length

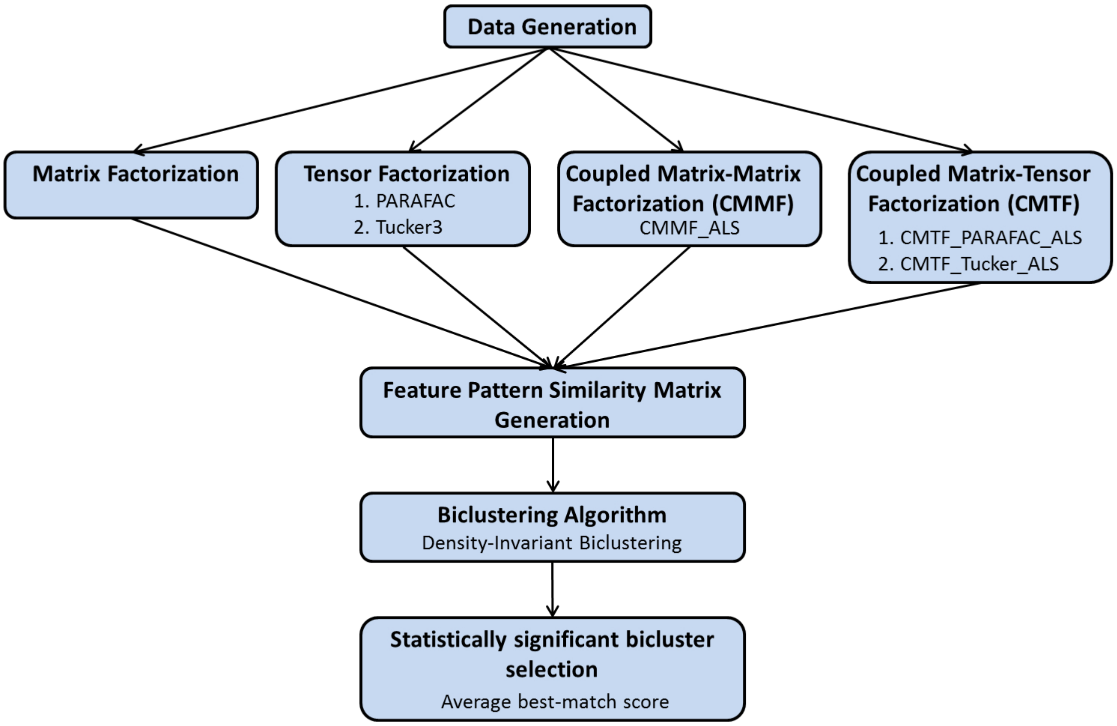

Data Generation
Matrix Factorization
Tensor Factorization
1. PARAFAC
2. Tucker3
Coupled Matrix-Matrix Factorization (CMMF)
CMMF_ALS
Coupled Matrix-Tensor Factorization (CMTF)
1. CMTF_PARAFAC_ALS
2. CMTF_Tucker_ALS
Feature Pattern Similarity Matrix Generation
Biclustering Algorithm
Density-Invariant Biclustering
Statistically significant bicluster selection
Average best-match score

Figure 2: Unified Biclustering Framework (UBF). In the first step, the data is generated as a matrix, a tensor, a coupled matrix-matrix, or a coupled matrix-tensor. In the second step, the data in various forms are factorized. In the third step, feature pattern similarity matrix is generated using the factor matrices of the decomposition. In the fourth step, we bicluster data points using density-invariant biclustering algorithm. In the final step, we find the most stable biclusters using average best-match score.

$L(C_i, C_j) > 3$, the proximity between two countries is set to zero. As a result, spatial proximity matrix reflects the likelihood of patients from two countries being involved in the same transmission event.

3.3 UBF: Unified Biclustering Framework

In order to analyze host-pathogen associations using various forms of the raw dataset, we propose the Unified Biclustering Framework (UBF). Based on this framework, we generate the data in the first step, which can be a matrix, a tensor, a coupled matrix-matrix, or a coupled matrix-tensor. In the second step, we decompose the dataset according to its form. In the third step, we generate the feature pattern similarity matrix. In the fourth step, we run the density-invariant biclustering (DIB) algorithm on the feature pattern similarity matrix. Finally, we find statistically significant biclusters and evaluate their biological relevance. Figure 2 shows the steps of UBF. The software for UBF is available at http://sourceforge.net/projects/ubf/. Next, we give the details of each step.

3.3.1 Data generation

The host-pathogen tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ can be coupled with genetic proximity matrix $\mathbf{Y} \in \mathbb{R}^{I \times M}$ and spatial proximity matrix $\mathbf{Z} \in \mathbb{R}^{J \times N}$. This flexibility leads to different data configurations which allows simultaneous factorization of different data blocks. Possible data configurations are shown in Figure 3. In data configuration

1, the host-pathogen tensor $\underline{\mathbf{X}}$ is summed and contracted along the time mode, and $\hat{\mathbf{X}} \in \mathbb{R}^{I \times J}$ is obtained and used without factorization. In data configuration 2, the original host-pathogen tensor $\underline{\mathbf{X}}$ is used. In data configuration 3, genetic proximity matrix $\mathbf{Y}$ is coupled with the host-pathogen tensor $\underline{\mathbf{X}}$ in the first mode, incorporating the genetic distance into domain knowledge. In data configuration 4, spatial proximity matrix $\mathbf{Z}$ is coupled with the host-pathogen tensor $\underline{\mathbf{X}}$ in the second mode, incorporating the spatial distance into domain knowledge. In data configuration 5, genetic proximity matrix $\mathbf{Y}$ and spatial proximity matrix $\mathbf{Z}$ are coupled with the host-pathogen tensor $\underline{\mathbf{X}}$ in the first and second mode respectively, incorporating both the genetic distance and spatial distance into domain knowledge. In data configuration 6, the host-pathogen tensor $\underline{\mathbf{X}}$ is contracted and summed along the time mode, keeping the genetic proximity matrix $\mathbf{Y}$ and spatial proximity matrix $\mathbf{Z}$ coupled with the contracted host-pathogen tensor, which is now the matrix $\hat{\mathbf{X}}$. We use all six configurations to find biclusters to associate spoligotypes and country of birth of tuberculosis patients, and test the effect of distance measures and time on detected groups.

3.3.2 Data factorization

In the second step of UBF, we factorize the dataset according to its form. If the data is a matrix, we use it as is. If it is a tensor, we use tensor decomposition methods, PARAFAC and Tucker3, and find the factor matrices for each mode. When the dataset is a coupled matrix-matrix or matrix-tensor, then we need to simultaneously factorize multiple matrices and/or tensors. We adopt the alternating least squares approach to solve coupled data factorizations. Next, we briefly outline the algorithms we use for coupled matrix-matrix factorization and coupled matrix-tensor factorization.

3.3.2.1 Coupled matrix-matrix factorization (CMMF): Coupled matrices are simultaneously factorized using the CMMF_ALS algorithm, which we outline next.

CMMF_ALS: The host-pathogen tensor contracted along the time mode becomes the matrix $\hat{\mathbf{X}} \in \mathbb{R}^{I \times J}$. Genetic proximity matrix $\mathbf{Y} \in \mathbb{R}^{I \times I}$ and spatial proximity matrix $\mathbf{Z} \in \mathbb{R}^{J \times J}$ are approximated as in the system of equations (1).

$$\begin{aligned} \hat{\mathbf{X}} &\approx \mathbf{A}\mathbf{B}' \\ \mathbf{Y} &\approx \mathbf{A}\mathbf{V}' \\ \mathbf{Z} &\approx \mathbf{B}\mathbf{W}' \,. \end{aligned} \tag{1}$$

We want to minimize the following loss function L_1, the sum of Frobenius norm of residuals for each data block:

$$L_1 = ||\hat{\mathbf{X}} - \mathbf{A}\mathbf{B}'||_F^2 + ||\mathbf{Y} - \mathbf{A}\mathbf{V}'||_F^2 + ||\mathbf{Z} - \mathbf{B}\mathbf{W}'||_F^2 \,. \tag{2}$$

To minimize L_1, we first initialize the factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{V}, \mathbf{W}$ using truncated SVD, and then alternately minimize the loss function by fixing one of them at a time.

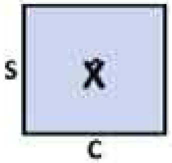
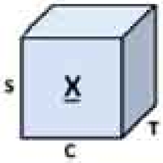
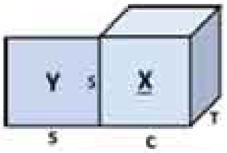
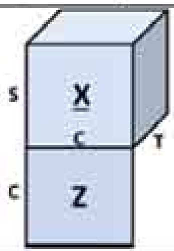
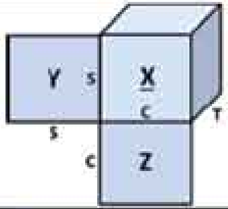
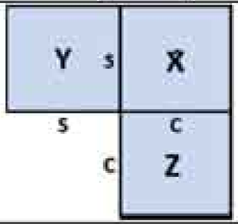
Number	Data configuration	Extra information	Method in UBF
1		—	MBF
2		Time	TBF
3		Time + genetic distance	CMTBF$_g$
4		Time + spatial distance	CMTBF$_s$
5		Time + genetic distance + spatial distance	CMTBF$_{gs}$
6		Genetic distance + spatial distance	CMMBF

Figure 3: Data configurations. The mode name S represents spoligotypes, C represents countries, and T represents time in years. The first configuration is a raw *Spoligotypes* × *Countries* **matrix decomposed using Matrix Biclustering Framework (MBF) as part of UBF. The second data configuration includes time information as the third mode of the tensor decomposed using Tensor Biclustering Framework (TBF) as part of UBF. Third, fourth and fifth data configurations are the results of concatenating the genetic proximity matrix, spatial proximity matrix, and both respectively, to the host-pathogen tensor. They are decomposed using Coupled Matrix-Tensor Biclustering Framework (CMTBF) as part of UBF. Finally, in data configuration 6, we exclude time information and decompose the resulting data using coupled matrix-matrix biclustering framework (CMMBF) as part of UBF.**

$$
\begin{aligned}
&\min_{A,B,V,W} ||\hat{\mathbf{X}} - \mathbf{A}\mathbf{B}'||_F^2 + ||\mathbf{Y} - \mathbf{A}\mathbf{V}'||_F^2 + ||\mathbf{Z} - \mathbf{B}\mathbf{W}'||_F^2 \\
&\min_{A,B,V,W} \mathrm{tr}\left(\left(\hat{\mathbf{X}} - \mathbf{A}\mathbf{B}'\right)\left(\hat{\mathbf{X}}' - \mathbf{B}\mathbf{A}'\right)\right) + \mathrm{tr}\left(\left(\mathbf{Y} - \mathbf{A}\mathbf{V}'\right)\left(\mathbf{Y}' - \mathbf{V}\mathbf{A}'\right)\right) \\
&\qquad + \mathrm{tr}\left(\left(\mathbf{Z} - \mathbf{B}\mathbf{W}'\right)\left(\mathbf{Z}' - \mathbf{W}\mathbf{B}'\right)\right) \\
&\min_{A,B,V,W} \mathrm{tr}\left(\hat{\mathbf{X}}\hat{\mathbf{X}}'\right) - 2\mathrm{tr}\left(\mathbf{B}\mathbf{A}'\hat{\mathbf{X}}\right) + \mathrm{tr}\left(\mathbf{A}\mathbf{B}'\mathbf{B}\mathbf{A}'\right) + \mathrm{tr}\left(\mathbf{Y}\mathbf{Y}'\right) - 2\mathrm{tr}\left(\mathbf{V}\mathbf{A}'\mathbf{Y}\right) + \\
&\qquad \mathrm{tr}\left(\mathbf{A}\mathbf{V}'\mathbf{V}\mathbf{A}'\right) + \mathrm{tr}\left(\mathbf{Z}\mathbf{Z}'\right) - 2\mathrm{tr}\left(\mathbf{W}\mathbf{B}'\mathbf{Z}\right) + \mathrm{tr}\left(\mathbf{B}\mathbf{W}'\mathbf{W}\mathbf{B}'\right) \\
&\min_{A,B,V,W} - 2\mathrm{tr}\left(\mathbf{B}\mathbf{A}'\hat{\mathbf{X}}\right) - 2\mathrm{tr}\left(\mathbf{V}\mathbf{A}'\mathbf{Y}\right) - 2\mathrm{tr}\left(\mathbf{W}\mathbf{B}'\mathbf{Z}\right) + \mathrm{tr}\left(\mathbf{A}\mathbf{B}'\mathbf{B}\mathbf{A}'\right) + \mathrm{tr}\left(\mathbf{A}\mathbf{V}'\mathbf{V}\mathbf{A}'\right) \\
&\qquad + \mathrm{tr}\left(\mathbf{B}\mathbf{W}'\mathbf{W}\mathbf{B}'\right) \qquad (3)
\end{aligned}
$$

Therefore, the objective function (3) is:

$$
\begin{aligned}
L = -2\mathrm{tr}\left(\mathbf{B}\mathbf{A}'\hat{\mathbf{X}}\right) - 2\mathrm{tr}\left(\mathbf{V}\mathbf{A}'\mathbf{Y}\right) - 2\mathrm{tr}\left(\mathbf{W}\mathbf{B}'\mathbf{Z}\right) + \mathrm{tr}\left(\mathbf{A}\mathbf{B}'\mathbf{B}\mathbf{A}'\right) + \mathrm{tr}\left(\mathbf{A}\mathbf{V}'\mathbf{V}\mathbf{A}'\right) \\
+\mathrm{tr}\left(\mathbf{B}\mathbf{W}'\mathbf{W}\mathbf{B}'\right) \ . \qquad (4)
\end{aligned}
$$

To minimize the loss function for $\mathbf{A}, \mathbf{B}, \mathbf{V}, \mathbf{W}$ after fixing other factor matrices, we take the derivative of objective function L in Equation (4), and set it to zero for each factor matrix, which gives the following update rules of matrices in CMMF_ALS:

Update for $\mathbf{A}$:

$$
\begin{aligned}
&\frac{\partial L}{\partial \mathbf{A}} = -2\hat{\mathbf{X}}\mathbf{B} - 2\mathbf{Y}\mathbf{V} + 2\mathbf{A}\mathbf{B}'\mathbf{B} + 2\mathbf{A}\mathbf{V}'\mathbf{V} = 0 \\
&\Longrightarrow \mathbf{A}\mathbf{B}'\mathbf{B} + \mathbf{A}\mathbf{V}'\mathbf{V} = \hat{\mathbf{X}}\mathbf{B} + \mathbf{Y}\mathbf{V} \\
&\qquad \mathbf{A} = \left(\hat{\mathbf{X}}\mathbf{B} + \mathbf{Y}\mathbf{V}\right) \setminus \left(\mathbf{B}'\mathbf{B} + \mathbf{V}'\mathbf{V}\right)
\end{aligned}
$$

Update for $\mathbf{B}$:

$$
\begin{aligned}
&\frac{\partial L}{\partial \mathbf{B}} = -2\hat{\mathbf{X}}'\mathbf{A} - 2\mathbf{Z}\mathbf{W} + 2\mathbf{B}\mathbf{A}'\mathbf{A} + 2\mathbf{B}\mathbf{W}'\mathbf{W} = 0 \\
&\Longrightarrow \mathbf{B}\mathbf{A}'\mathbf{A} + \mathbf{B}\mathbf{W}'\mathbf{W} = \hat{\mathbf{X}}'\mathbf{A} + \mathbf{Z}\mathbf{W} \\
&\qquad \mathbf{B} = \left(\hat{\mathbf{X}}'\mathbf{A} + \mathbf{Z}\mathbf{W}\right) \setminus \left(\mathbf{A}'\mathbf{A} + \mathbf{W}'\mathbf{W}\right)
\end{aligned}
$$

Update for $\mathbf{V}$:

$$
\begin{aligned}
&\frac{\partial L}{\partial \mathbf{V}} = -2\mathbf{Y}'\mathbf{A} + 2\mathbf{V}\mathbf{A}'\mathbf{A} = 0 \\
&\Longrightarrow \mathbf{V}\mathbf{A}'\mathbf{A} = \mathbf{Y}'\mathbf{A} \\
&\qquad \mathbf{V} = \left(\mathbf{Y}'\mathbf{A}\right) \setminus \left(\mathbf{A}'\mathbf{A}\right)
\end{aligned}
$$

Update for $\mathbf{W}$:

$$
\begin{aligned}
&\frac{\partial L}{\partial \mathbf{W}} = -2\mathbf{Z}'\mathbf{B} + 2\mathbf{W}\mathbf{B}'\mathbf{B} = 0 \\
&\Longrightarrow \mathbf{W}\mathbf{B}'\mathbf{B} = \mathbf{Z}'\mathbf{B} \\
&\qquad \mathbf{W} = \left(\mathbf{Z}'\mathbf{B}\right) \setminus \left(\mathbf{B}'\mathbf{B}\right)
\end{aligned}
$$

where \ represents right matrix division. The complete CMMF_ALS procedure is summarized in Algorithm 1. In this algorithm, the function `svd_mmf`($\hat{\mathbf{X}}$, $\mathbf{Y}$, $\mathbf{Z}$) initializes the factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{V}, \mathbf{W}$ using truncated SVD with $min(J, M, N)$ components.

Algorithm 1 `CMMF_ALS`($\hat{\mathbf{X}} \in \mathbb{R}^{I \times J}$,$\mathbf{Y} \in \mathbb{R}^{I \times M}$,$\mathbf{Z} \in \mathbb{R}^{J \times N}$)

```
[A, B, V, W] = svd_mmf($\hat{\mathbf{X}}$, $\mathbf{Y}$, $\mathbf{Z}$)
loss(current) = $||\hat{\mathbf{X}} - \mathbf{AB}'||_F^2 + ||\mathbf{Y} - \mathbf{AV}'||_F^2 + ||\mathbf{Z} - \mathbf{BW}'||_F^2$
loss(prev) = loss(current)
$count = 0$
while $((count == 0) \;||\; (0 < count \leq 10^3 \;\&\&\; \frac{|loss(current) - loss(prev)|}{loss(prev)} > 10^{-8}))$ do
    $count + +$
    // Solve for A
    $\mathbf{A} = \left(\hat{\mathbf{X}}\mathbf{B} + \mathbf{YV}\right) \setminus (\mathbf{B}'\mathbf{B} + \mathbf{V}'\mathbf{V})$
    // Solve for B
    $\mathbf{B} = \left(\hat{\mathbf{X}}'\mathbf{A} + \mathbf{ZW}\right) \setminus (\mathbf{A}'\mathbf{A} + \mathbf{W}'\mathbf{W})$
    // Solve for V
    $\mathbf{V} = (\mathbf{Y}'\mathbf{A}) \setminus (\mathbf{A}'\mathbf{A})$
    // Solve for W
    $\mathbf{W} = (\mathbf{Z}'\mathbf{B}) \setminus (\mathbf{B}'\mathbf{B})$
    loss(prev) = loss(current)
    loss(current) = $||\hat{\mathbf{X}} - \mathbf{AB}'||_F^2 + ||\mathbf{Y} - \mathbf{AV}'||_F^2 + ||\mathbf{Z} - \mathbf{BW}'||_F^2$
end while
```

3.3.2.2 Coupled matrix-tensor factorization (CMTF): Coupled matrices and tensors can be simultaneously factorized. For this purpose, we used modifications of PARAFAC and Tucker3 methods. CMTF_PARAFAC_ALS decomposes the tensor using PARAFAC while factorizing the coupled matrices simultaneously. CMTF_PARAFAC_ALS algorithm and its variations exist in the literature. We built another algorithm, extension of Tucker3 to coupled matrix-tensor factorization. CMTF_Tucker_ALS algorithm decomposes the tensor using Tucker3, while simultaneously factorizing the coupled matrices. In the next section, we give the details of these algorithms.

CMTF_PARAFAC_ALS: Given the host-pathogen tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ coupled with genetic proximity matrix $\mathbf{Y} \in \mathbb{R}^{I \times I}$ and spatial proximity matrix $\mathbf{Z} \in \mathbb{R}^{J \times J}$, we approximate them as follows:

$$\begin{aligned} \mathbf{X_{(1)}} &\approx \mathbf{A}\left(\mathbf{C} \odot \mathbf{B}\right)' \\ \mathbf{Y} &\approx \mathbf{AV}' \\ \mathbf{Z} &\approx \mathbf{BW}' \end{aligned} \tag{5}$$

where $\odot$ denotes the Khatri-Rao product. We want to minimize the following loss function which is the sum of squared Frobenius norm of residuals for each data block:

$$L_2 = ||\mathbf{X_{(1)}} - \mathbf{A}\left(\mathbf{C} \odot \mathbf{B}\right)'||_F^2 + ||\mathbf{Y} - \mathbf{AV}'||_F^2 + ||\mathbf{Z} - \mathbf{BW}'||_F^2 \,. \tag{6}$$

CMTF_PARAFAC_ALS is also known as CMTF_ALS algorithm in the literature, which is detailed in earlier studies [26]. Therefore, we skip the details of the algorithm, and only focus on the update step for each factor matrix. Minimization for $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{V}, \mathbf{W}$ alternately returns the following updates at each step of CMTF_PARAFAC_ALS:

Update for $\mathbf{A}$:

$$\begin{aligned}
&\min_{A} ||\mathbf{X_{(1)}} - \mathbf{A}\left(\mathbf{C} \odot \mathbf{B}\right)'||_F^2 + ||\mathbf{Y} - \mathbf{A}\mathbf{V}'||_F^2 \\
&\min_{A} ||\underbrace{\left[\mathbf{X_{(1)}}\ \ \mathbf{Y}\right]}_{T} - \mathbf{A}\underbrace{\left[\left(\mathbf{C} \odot \mathbf{B}\right)'\ \ \mathbf{V}'\right]}_{K}||_F^2 \\
&\Longrightarrow \mathbf{A} = \left(\mathbf{T}\mathbf{K}'\right) / \left(\mathbf{K}\mathbf{K}'\right)
\end{aligned}$$

Update for $\mathbf{B}$:

$$\begin{aligned}
&\min_{B} ||\mathbf{X_{(2)}} - \mathbf{B}\left(\mathbf{C} \odot \mathbf{A}\right)'||_F^2 + ||\mathbf{Z} - \mathbf{B}\mathbf{W}'||_F^2 \\
&\min_{B} ||\underbrace{\left[\mathbf{X_{(2)}}\ \ \mathbf{Z}\right]}_{T} - \mathbf{B}\underbrace{\left[\left(\mathbf{C} \odot \mathbf{A}\right)'\ \ \mathbf{W}'\right]}_{K}||_F^2 \\
&\Longrightarrow \mathbf{B} = \left(\mathbf{T}\mathbf{K}'\right) / \left(\mathbf{K}\mathbf{K}'\right)
\end{aligned}$$

Update for $\mathbf{C}$:

$$\begin{aligned}
&\min_{C} ||\underbrace{\mathbf{X_{(3)}}}_{T} - \mathbf{C}\underbrace{\left(\mathbf{B} \odot \mathbf{A}\right)'}_{K}||_F^2 \\
&\Longrightarrow \mathbf{C} = \left(\mathbf{T}\mathbf{K}'\right) / \left(\mathbf{K}\mathbf{K}'\right)
\end{aligned}$$

Update for $\mathbf{V}$:

$$\begin{aligned}
&\min_{V} ||\mathbf{Y} - \mathbf{A}\mathbf{V}'||_F^2 \\
&\Longrightarrow \mathbf{V} = \left(\left(\mathbf{A}'\mathbf{A}\right) \setminus \left(\mathbf{A}'\mathbf{Y}\right)\right)'
\end{aligned}$$

Update for $\mathbf{W}$:

$$\begin{aligned}
&\min_{W} ||\mathbf{Z} - \mathbf{B}\mathbf{W}'||_F^2 \\
&\Longrightarrow \mathbf{W} = \left(\left(\mathbf{B}'\mathbf{B}\right) \setminus \left(\mathbf{B}'\mathbf{Z}\right)\right)'
\end{aligned}$$

CMTF_Tucker_ALS: Next, we extend Tucker3 method to CMTF_Tucker_ALS for coupled matrix-tensor decomposition. This algorithm comes with the flexibility of factorizing the tensor using different number of components for each mode, while simultaneously factorizing the coupled matrices. The host-pathogen tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$, genetic proximity matrix $\mathbf{Y} \in \mathbb{R}^{I \times I}$ and spatial proximity matrix $\mathbf{Z} \in \mathbb{R}^{J \times J}$ are approximated as in the system of equations (7).

$$\begin{aligned}
\mathbf{X_{(1)}} &\approx \mathbf{A}\mathbf{G_{(1)}}\left(\mathbf{C}' \otimes \mathbf{B}'\right) \\
\mathbf{Y} &\approx \mathbf{A}\mathbf{V}' \\
\mathbf{Z} &\approx \mathbf{B}\mathbf{W}'
\end{aligned} \tag{7}$$

where $\otimes$ denotes the Kronecker product. Note that in the Tucker3 model, the factor matrices $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$, $\mathbf{C} \in \mathbb{R}^{K \times R}$ are orthogonal. Then, tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ can be decomposed using a (P, Q, R)-component Tucker3 model, while simultaneously factorizing $\mathbf{Y} \in \mathbb{R}^{I \times I}$ and $\mathbf{Z} \in \mathbb{R}^{J \times J}$ with the factor matrices of the shared mode. We want to minimize the loss function L_3 in Equation (8), which is the sum of squared Frobenius norm of residuals for each data block.

$$L_3 = ||\mathbf{X}_{(\mathbf{1})} - \mathbf{A}\mathbf{G}_{(\mathbf{1})}\left(\mathbf{C}' \otimes \mathbf{B}'\right)||_F^2 + ||\mathbf{Y} - \mathbf{A}\mathbf{V}'||_F^2 + ||\mathbf{Z} - \mathbf{B}\mathbf{W}'||_F^2 \,. \tag{8}$$

To minimize L_3, we first initialize the factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{V}, \mathbf{W}$ using truncated SVD, and then alternately minimize the loss function for one of the variables at a time, while fixing the other variables. The following steps in Equation (9) reformulate the minimization of the loss function.

$$\begin{aligned}
&\min_A ||\mathbf{X}_{(\mathbf{1})} - \mathbf{A}\mathbf{G}_{(\mathbf{1})}\left(\mathbf{C}' \otimes \mathbf{B}'\right)||_F^2 + ||\mathbf{Y} - \mathbf{A}\mathbf{V}'||_F^2 \\
&\min_A || \left[\mathbf{X}_{(\mathbf{1})}\ \mathbf{Y}\right] - \left[\mathbf{A}\mathbf{G}_{(\mathbf{1})}\left(\mathbf{C}' \otimes \mathbf{B}'\right)\ \mathbf{A}\mathbf{V}'\right] ||_F^2 \\
&\min_A || \left[\mathbf{X}_{(\mathbf{1})}\ \mathbf{Y}\right] - \left[\mathbf{A}\mathbf{A}'\underbrace{\mathbf{X}_{(\mathbf{1})}\left(\mathbf{C}\mathbf{C}' \otimes \mathbf{B}\mathbf{B}'\right)}_{\mathbf{M}_1}\ \mathbf{A}\mathbf{V}'\right] ||_F^2 \\
&\min_A || \left[\mathbf{X}_{(\mathbf{1})}\ \mathbf{Y}\right] - \left[\mathbf{A}\mathbf{A}'\mathbf{M}_1\ \mathbf{A}\mathbf{V}'\right] ||_F^2 \\
&\min_A + \operatorname{tr}\left(\left(\left[\mathbf{X}_{(\mathbf{1})}\ \mathbf{Y}\right] - \left[\mathbf{A}\mathbf{A}'\mathbf{M}_1\ \mathbf{A}\mathbf{V}'\right]\right)\left(\left[\mathbf{X}_{(\mathbf{1})}\ \mathbf{Y}\right]' - \left[\mathbf{A}\mathbf{A}'\mathbf{M}_1\ \mathbf{A}\mathbf{V}'\right]'\right)\right) \\
&\min_A + \operatorname{tr}\left(\left[\mathbf{X}_{(\mathbf{1})}\ \mathbf{Y}\right]\left[\mathbf{X}_{(\mathbf{1})}\ \mathbf{Y}\right]'\right) - 2\operatorname{tr}\left(\left[\mathbf{X}_{(\mathbf{1})}\ \mathbf{Y}\right]\left[\mathbf{M}_1'\mathbf{A}\mathbf{A}'\ ;\ \mathbf{V}\mathbf{A}'\right]\right) \\
&\qquad + \operatorname{tr}\left(\left[\mathbf{A}\mathbf{A}'\mathbf{M}_1\ \mathbf{A}\mathbf{V}'\right]\left[\mathbf{M}_1\mathbf{A}\mathbf{A}'\ ;\ \mathbf{V}\mathbf{A}'\right]\right) \\
&\min_A - 2\operatorname{tr}\left(\left[\mathbf{X}_{(\mathbf{1})}\ \mathbf{Y}\right]\left[\mathbf{M}_1'\mathbf{A}\mathbf{A}'\ ;\ \mathbf{V}\mathbf{A}'\right]\right) + \operatorname{tr}\left(\mathbf{A}\mathbf{A}'\mathbf{M}_1\mathbf{M}_1'\mathbf{A}\mathbf{A}' + \mathbf{A}\mathbf{V}'\mathbf{V}\mathbf{A}'\right) \\
&\min_A - 2\operatorname{tr}\left(\mathbf{X}_{(\mathbf{1})}\mathbf{M}_1'\mathbf{A}\mathbf{A}' + \mathbf{Y}\mathbf{V}\mathbf{A}'\right) + \operatorname{tr}\left(\mathbf{A}\mathbf{A}'\mathbf{M}_1\mathbf{M}_1'\mathbf{A}\mathbf{A}' + \mathbf{A}\mathbf{V}'\mathbf{V}\mathbf{A}'\right) \\
&\min_A - 2\operatorname{tr}\left(\mathbf{X}_{(\mathbf{1})}\mathbf{M}_1'\mathbf{A}\mathbf{A}'\right) - 2\operatorname{tr}\left(\mathbf{Y}\mathbf{V}\mathbf{A}'\right) + \operatorname{tr}\left(\mathbf{A}\mathbf{A}'\mathbf{M}_1\mathbf{M}_1'\mathbf{A}\mathbf{A}'\right) + \operatorname{tr}\left(\mathbf{A}\mathbf{V}'\mathbf{V}\mathbf{A}'\right) \\
&\min_A - 2\operatorname{tr}\left(\mathbf{M}_1\mathbf{M}_1'\mathbf{A}\mathbf{A}'\right) - 2\operatorname{tr}\left(\mathbf{Y}\mathbf{Y}'\mathbf{A}\mathbf{A}'\right) + \operatorname{tr}\left(\mathbf{A}'\mathbf{M}_1\mathbf{M}_1'\mathbf{A}\right) + \operatorname{tr}\left(\mathbf{A}\mathbf{A}'\mathbf{Y}\mathbf{Y}'\mathbf{A}\mathbf{A}'\right) \\
&\min_A - 2\operatorname{tr}\left(\mathbf{A}'\mathbf{M}_1\mathbf{M}_1'\mathbf{A}\right) - 2\operatorname{tr}\left(\mathbf{A}'\mathbf{Y}\mathbf{Y}'\mathbf{A}\right) + \operatorname{tr}\left(\mathbf{A}'\mathbf{M}_1\mathbf{M}_1'\mathbf{A}\right) + \operatorname{tr}\left(\mathbf{A}'\mathbf{Y}\mathbf{Y}'\mathbf{A}\right) \\
&\min_A - \operatorname{tr}\left(\mathbf{A}'\mathbf{M}_1\mathbf{M}_1'\mathbf{A}\right) - \operatorname{tr}\left(\mathbf{A}'\mathbf{Y}\mathbf{Y}'\mathbf{A}\right) \qquad (9) \\
&\text{s.t. } \ \mathbf{A}'\mathbf{A} = \mathbf{I}
\end{aligned}$$

where $\mathbf{M}_1 = \mathbf{X}_{(\mathbf{1})}\left(\mathbf{C}\mathbf{C}' \otimes \mathbf{B}\mathbf{B}'\right)$. The Lagrangian of this function is:

$$L_A = -\operatorname{tr}\left(\mathbf{A}'\mathbf{M}_1\mathbf{M}_1'\mathbf{A}\right) - \operatorname{tr}\left(\mathbf{A}'\mathbf{Y}\mathbf{Y}'\mathbf{A}\right) + \operatorname{tr}\left(\lambda\left(\mathbf{A}'\mathbf{A} - \mathbf{I}\right)\right)$$

where λ are the Lagrangian multipliers for the orthogonality constraint $\mathbf{A}'\mathbf{A} = \mathbf{I}$. The derivative of L_A with respect to $\mathbf{A}$ set to zero returns the following equation:

$$\begin{aligned}
&\frac{\partial L_A}{\partial \mathbf{A}} = -2\mathbf{M}_1\mathbf{M}_1'\mathbf{A} - 2\mathbf{Y}\mathbf{Y}'\mathbf{A} + \lambda\left(2\mathbf{A}\right) = 0 \\
&\Longrightarrow \left(\mathbf{M}_1\mathbf{M}_1' + \mathbf{Y}\mathbf{Y}'\right)\mathbf{A} = \lambda\mathbf{A} \qquad (10)
\end{aligned}$$

The optimal solution of (9) must satisfy Equation (10). Thus, $\mathbf{A}$ is composed of first P largest eigenvectors of $\left(\mathbf{M}_1\mathbf{M}_1' + \mathbf{Y}\mathbf{Y}'\right)$. We denote it as follows:

$$\mathbf{A} = \mathrm{EVD}\left(\mathbf{M}_1\mathbf{M}_1' + \mathbf{YY}', P\right) \ . \tag{11}$$

Similarly, for the second mode, we write the loss function L_3 in Equation (8) by matricizing the tensor along the second mode. Then, the objective function is:

$$\begin{aligned} &\min_{B} - \mathrm{tr}\left(\mathbf{B}'\mathbf{M}_2\mathbf{M}_2'\mathbf{B}\right) - \mathrm{tr}\left(\mathbf{B}'\mathbf{ZZ}'\mathbf{B}\right) \\ &\text{s.t. } \ \mathbf{B}'\mathbf{B} = \mathbf{I} \end{aligned} \tag{12}$$

where $\mathbf{M_2} = \mathbf{X_{(2)}}\left(\mathbf{CC}' \otimes \mathbf{AA}'\right)$. The Lagrangian of this objective function is:

$$L_B = -\mathrm{tr}\left(\mathbf{B}'\mathbf{M}_2\mathbf{M}_2'\mathbf{B}\right) - \mathrm{tr}\left(\mathbf{B}'\mathbf{ZZ}'\mathbf{B}\right) + \mathrm{tr}\left(\lambda\left(\mathbf{B}'\mathbf{B} - \mathbf{I}\right)\right)$$

where λ are the Lagrangian multipliers for the orthogonality constraint $\mathbf{B}'\mathbf{B} = \mathbf{I}$. The derivative of L_B with respect to $\mathbf{B}$ set to zero returns the following equation:

$$\begin{aligned} &\frac{\partial L_B}{\partial \mathbf{B}} = -2\mathbf{M}_2\mathbf{M}_2'\mathbf{B} - 2\mathbf{ZZ}'\mathbf{B} + \lambda\left(2\mathbf{B}\right) = 0 \\ &\Longrightarrow \left(\mathbf{M}_2\mathbf{M}_2' + \mathbf{ZZ}'\right)\mathbf{B} = \lambda\mathbf{B} \end{aligned}$$

which means that $\mathbf{B}$ is composed of first Q largest eigenvectors of $(\mathbf{M}_2\mathbf{M}_2' + \mathbf{ZZ}')$. We denote it as follows:

$$\mathbf{B} = \mathrm{EVD}\left(\mathbf{M_2}\mathbf{M_2}' + \mathbf{ZZ}', Q\right) \ . \tag{13}$$

For the uncoupled third mode, we write the objective function L_3 in Equation (8) by matricizing the tensor along the third mode. The objective function is as follows:

$$\begin{aligned} &\min_{C} - \mathrm{tr}\left(\mathbf{C}'\mathbf{M}_3\mathbf{M}_3'\mathbf{C}\right) \\ &\text{s.t. } \ \mathbf{C}'\mathbf{C} = \mathbf{I} \end{aligned} \tag{14}$$

where $\mathbf{M_3} = \mathbf{X_{(3)}}\left(\mathbf{BB}' \otimes \mathbf{AA}'\right)$. The Lagrangian of this function is:

$$L_C = -\mathrm{tr}\left(\mathbf{C}'\mathbf{M}_3\mathbf{M}_3'\mathbf{C}\right) + \lambda\left(\mathrm{tr}\left(\mathbf{C}'\mathbf{C} - \mathbf{I}\right)\right) \ .$$

The derivative of L_C with respect to $\mathbf{C}$ set to zero returns the following equation:

$$\begin{aligned} &\frac{\partial L_C}{\partial \mathbf{C}} = -2\mathbf{M}_3\mathbf{M}_3'\mathbf{C} + \lambda\left(2\mathbf{C}\right) = 0 \\ &\Longrightarrow \mathbf{M}_3\mathbf{M}_3'\mathbf{C} = \lambda\mathbf{C} \end{aligned}$$

which means that $\mathbf{C}$ is composed of first R largest eigenvectors of $\mathbf{M}_3\mathbf{M}_3'$, or equivalently, first R left singular vectors of $\mathbf{M}_3$. We denote it as follows:

$$\mathbf{C} = \mathrm{SVD}\left(\mathbf{M_3}, R\right)\ . \tag{15}$$

The complete CMTF_Tucker_ALS procedure using these update rules is summarized in Algorithm 2. Note that the function call `hosvd_Tucker(`$\underline{\mathbf{X}}$, $[P, Q, R]$`)` at the beginning of the algorithm initializes factor matrices via truncated SVD using P, Q, R components respectively for each mode. The function `unfoldall(`$\underline{\mathbf{X}}$`)` matricizes the tensor along each mode.

Algorithm 2 `CMTF_Tucker_ALS(`$\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$`,`$\mathbf{Y} \in \mathbb{R}^{I \times M}$`,`$\mathbf{Z} \in \mathbb{R}^{J \times N}$`, [`$P, Q, R$`])`

```
$[\mathbf{A}, \mathbf{B}, \mathbf{C}, \underline{\mathbf{G}}] = \texttt{hosvd\_Tucker}(\underline{\mathbf{X}}, [P, Q, R]);$
$\mathbf{V} = ((\mathbf{A'A}) \backslash (\mathbf{A'Y}))'$
$\mathbf{W} = ((\mathbf{B'B}) \backslash (\mathbf{B'Z}))'$
$[\mathbf{X_{(1)}}, \mathbf{X_{(2)}}, \mathbf{X_{(3)}}] = \texttt{unfoldall}(\underline{\mathbf{X}})$
$[\mathbf{G_{(1)}}, \mathbf{G_{(2)}}, \mathbf{G_{(3)}}] = \texttt{unfoldall}(\underline{\mathbf{G}})$
$\text{loss(current)} = ||\mathbf{X_{(1)}} - \mathbf{AG_{(1)}}\left(\mathbf{C'} \otimes \mathbf{B'}\right)||_F^2 + ||\mathbf{Y} - \mathbf{AV'}||_F^2 + ||\mathbf{Z} - \mathbf{BW'}||_F^2$
$\text{loss(prev)} = \text{loss(current)}$
$count = 0$
while $((count == 0)\ ||\ (0 < count \leq 10^3\ \&\&\ \frac{|loss(current) - loss(prev)|}{loss(prev)} > 10^{-8}))$ do
    $count++$
    // Solve for A
    $\mathbf{M_1} = \mathbf{X_{(1)}}\left(\mathbf{CC'} \otimes \mathbf{BB'}\right)$
    $\mathbf{A} = \mathrm{EVD}\left(\mathbf{M_1 M_1}' + \mathbf{YY'}, P\right)$
    // Solve for B
    $\mathbf{M_2} = \mathbf{X_{(2)}}\left(\mathbf{CC'} \otimes \mathbf{AA'}\right)$
    $\mathbf{B} = \mathrm{EVD}\left(\mathbf{M_2 M_2}' + \mathbf{ZZ'}, Q\right)$
    // Solve for C
    $\mathbf{M_3} = \mathbf{X_{(3)}}\left(\mathbf{BB'} \otimes \mathbf{AA'}\right)$
    $\mathbf{C} = \mathrm{SVD}\left(\mathbf{M_3}, R\right)$
    // Solve for V
    $\mathbf{V} = ((\mathbf{A'A}) \backslash (\mathbf{A'Y}))'$
    // Solve for W
    $\mathbf{W} = ((\mathbf{B'B}) \backslash (\mathbf{B'Z}))'$
    $\text{loss(prev)} = \text{loss(current)}$
    $\text{loss(current)} = ||\mathbf{X_{(1)}} - \mathbf{AG_{(1)}}\left(\mathbf{C'} \otimes \mathbf{B'}\right)||_F^2 + ||\mathbf{Y} - \mathbf{AV'}||_F^2 + ||\mathbf{Z} - \mathbf{BW'}||_F^2$
end while
```

3.3.3 Feature pattern similarity matrix generation

We calculate the similarity of feature patterns of a spoligotype s and country c by calculating cosine similarity between feature pattern vectors of them. This is calculated in different ways for different forms of input data. If the input data is a matrix, then the matrix itself is used as the feature pattern similarity matrix (FPSM). If the data is in tensor form, then FPSM is calculated for PARAFAC as follows. Assume that R-component PARAFAC model on the data matrix returns factor matrix $\mathbf{A} \in \mathbb{R}^{I \times R}$ for the first mode and factor matrix $\mathbf{B} \in \mathbb{R}^{J \times R}$ for the second mode. Then, we first normalize the rows of $\mathbf{A}$ and $\mathbf{B}$, and calculate the feature pattern similarity matrix $FPSM$ as follows:

$$\mathrm{FPSM}_{ij} = \begin{cases} \dfrac{\mathbf{A}_{i.}\ \mathbf{B}'_{j.}}{||\mathbf{A}_{i.}||\ ||\mathbf{B}_{j.}||}, & \text{if } N(i,j) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

where $N(i,j)$ represents the number of patients from country j infected with strain i, and $\mathbf{A}_{i.}$ represents the i-th row of $\mathbf{A}$. This calculation is equivalent to cosine similarity of feature vector of i-th sample of $\mathbf{A}$ and feature vector of j-th sample of $\mathbf{B}$, only if there is at least one patient from country j infected with strain i. Calculation of feature pattern matrix after applying Tucker3 model is slightly different. Assume that (P,Q,R)-component Tucker3 model on the data matrix returns factor matrix $\mathbf{A} \in \mathbb{R}^{I \times P}$ for the first mode, factor matrix $\mathbf{B} \in \mathbb{R}^{J \times Q}$ for the second mode, and the core tensor $\underline{\mathbf{G}} \in \mathbb{R}^{P \times Q \times R}$. First, we contract and sum the core tensor $\underline{\mathbf{G}}$ along the third mode and obtain $\hat{\mathbf{G}}$ matrix to calculate the level of interaction between the factors of $\mathbf{A}$ and $\mathbf{B}$. We normalize the rows of $\mathbf{A}\hat{\mathbf{G}}$ and $\mathbf{B}$. Finally, we calculate the feature pattern similarity matrix as the cosine similarity of $\mathbf{A}\hat{\mathbf{G}}$ and $\mathbf{B}$, in Equation (17).

$$\begin{aligned} \hat{\mathbf{G}}_{pq} &= \sum_{r=1}^{R} \underline{\mathbf{G}}_{pqr} \\ \mathrm{FPSM}_{ij} &= \begin{cases} \dfrac{\mathbf{A}_{i.}\,\hat{\mathbf{G}}}{||\mathbf{A}_{i.}\,\hat{\mathbf{G}}||}\,\dfrac{\mathbf{B}'_{j.}}{||\mathbf{B}_{j.}||}, & \text{if } N(i,j) > 0 \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{17}$$

For coupled factorizations, we use the same equations. After coupled matrix-matrix decomposition, we use Equation (16) to find the feature pattern similarity matrix. For coupled matrix-tensor factorization, if CMTF_PARAFAC_ALS is used for factorization, then FPSM is calculated using Equation (16). If CMTF_Tucker_ALS is used for factorization, then Equation (17) is used to calculate FPSM.

3.3.4 Density-invariant biclustering

In this section, we introduce a novel biclustering algorithm based on an existing algorithm and several graph attributes. First, we discretize the input matrix and use it as input to BiMax algorithm to find inclusion-maximal biclusters [19]. Then, we use these biclusters as seed, and find density and variance of these biclusters, which are bicliques. Finally, we find the density-invariant biclusters among candidate inclusion-maximal biclusters.

Given the feature pattern similarity matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$, we use density-invariant biclustering to find coherent biclusters. Let $G = (U, V, E)$ represent a bipartite graph, where U represents the set of genes, or rows in $\mathbf{X}$, V represents the set of conditions, or columns in $\mathbf{X}$, and E represents the weight of the edges connecting vertex set U and vertex set V. The weights E are equivalent to values of matrix $\mathbf{X}$. We want to find biclusters of the following form:

$$B_i = (U_i, V_i, E_i) \tag{18}$$

where $B = \bigcup_{i=1}^{n} B_i$ is a biclustering of rows and columns of $\mathbf{X}$. Each bicluster associates a set of rows, in this case spoligotypes, to a set of columns, in this case countries. Notice that each bicluster maps to a submatrix of the original data matrix.

Density-invariant biclustering algorithm first discretizes edge weights using a weight threshold th, and converts the input matrix into a binary matrix $\mathbf{D}$. Then we use the binary inclusion-maximal biclustering algorithm (BiMax) by Prelic et al. on this binary matrix and find a set of candidate biclusters [19]. These biclusters are inclusion-maximal, because the submatrices corresponding to these biclusters are all 1's, and there is no other bicluster which is a superset of it. Output of BiMax algorithm after discretization returns a good starting point for density-invariant biclustering algorithm. Next, we focus on these candidate biclusters. For this purpose, we define the density and variance of a graph.

Definition 1. ***Density of a graph:*** *Density of a graph is the average weight of its edges. Given a graph $G = (V, E)$ where w(e) represents the weight of edge $e \in E$, the density of graph G is calculated as follows:*

$$d(G) = \frac{\sum\limits_{e \in E} w(e)}{\binom{|V|}{2}} .$$

Definition 2. ***Variance of a graph:*** *Variance of a graph is the standard deviation of its edge weights. Given a graph $G = (V, E)$ where w(e) represents the weight of edge $e \in E$, the variance of graph G is calculated as follows:*

$$v(G) = \sqrt{\frac{1}{|E| - 1} \sum_{e \in E} \left(w\left(e\right) - \bar{w}\right)^2} .$$

Using the density and variance of graphs, we can define a new set of graphs which are bounded by their edge weights. Next, we define the α-dense β-variant biclusters, or density-invariant biclusters, which are graphs of the form $B = (U, V, E)$ with density $d(B) \geq \alpha$ and variance $v(B) \leq \beta$, and similarly for all one-vertex-induced subgraphs of $B = (U, V, E)$.

Definition 3. ***Density-invariant bicluster:*** *Let $B = (U, V, E)$ be a bicluster, where edges in E connect vertices in U to vertices in V. Bicluster B is an α-dense bicluster if $d(B) \geq \alpha$, and it is a β-variant bicluster if $v(B) \leq \beta$. Define B' as an induced subgraph of B after removing one vertex, either from vertex set U or vertex set V. Bicluster $B = (U, V, E)$ is an (α, β)-density-invariant bicluster, or density-invariant bicluster, if B and all its one-vertex-induced subgraphs are α-dense β-variant. In short, bicluster B is a density-invariant bicluster if the following conditions hold:*

1. $d(B) \geq \alpha$, $v(B) \leq \beta$
2. $d(B') \geq \alpha$, $v(B') \leq \beta \quad \forall B' = B \setminus \{m\}$ *where* $m \in U \cup V$, $|B'| > 0$.

Notice that a density-invariant bicluster forms a biclique with average weight bounded from below, and variance of weights bounded from above. All induced subgraphs obtained after removing one vertex from a density-invariant bicluster are still α-dense and β-variant, but not necessarily density-invariant biclusters. At this point, we define strong antimonotonicity of a graph, which was introduced in Pao et al. [27].

Definition 4. ***Strong antimonotonicity:*** *A graph attribute is strong antimonotone if for each graph $G = (V, E)$ with the attribute, every induced subgraph $G' = G - \{v\}$ has the attribute, where $v \in V$.*

According to the definition of strong antimonotonicity, the attribute of being a density-invariant graph or bicluster is not strongly antimonotone. This is because the vertex-induced subgraphs of the original graph are α-dense and β-variant, but their vertex-induced subgraphs need not be (α, β)-density-invariant biclusters.

Finally, we iterate over candidate biclusters found as output from BiMax algorithm and find density-invariant biclusters among these candidate biclusters. This results in strongly connected and more homogeneous biclusters. Algorithm 3 summarizes `DensityInvariantBiclustering` procedure.

In `DensityInvariantBiclustering` algorithm, `discretize`$(\mathbf{X}, th)$ function discretizes the input data matrix as follows:

$$\mathbf{D}_{ij} = \begin{cases} 1, & \text{if } \mathbf{X}_{ij} \geq th \\ 0, & \text{otherwise.} \end{cases}$$

`BiMax` algorithm is run on this binary matrix $\mathbf{D}$, and inclusion-maximal biclusters are obtained. Then, among these candidate biclusters, density-invariant biclusters are found.

Algorithm 3 `Biclusters = DensityInvariantBiclustering(`$\mathbf{X} \in \mathbb{R}^{I\times J}$, th, α, β`)`

```
Input: Data matrix $\mathbf{X} \in \mathbb{R}^{I\times J}$, discretization threshold $th$, density threshold $\alpha$, variance threshold $\beta$.
Output: Density-invariant biclusters Biclusters.
D = discretize($\mathbf{X}, th$)
CandidateBiclusters = BiMax($\mathbf{D}$)
Biclusters = $\emptyset$
for i=1:1:length(CandidateBiclusters) do
    $B(U,V,E)$ = CandidateBiclusters(i)
    $check1 = (d(B) \geq \alpha)$ && $(v(B) \leq \beta)$
    $check2 = true$
    M = $U \cup V$
    for j=1:1:length(M) do
        $m = M(j)$
        $B' = B \setminus \{m\}$
        if $((B' \neq \emptyset)$ && $!(d(B') \geq \alpha$ && $v(B') \leq \beta))$ then
            $check2 = false$
            break
        end if
    end for
    if ($check1$ && $check2$) then
        Biclusters = Biclusters $\cup\{B\}$
    end if
end for
```

3.3.5 Statistically significant bicluster selection

In order to find statistically significant biclusters, we sample 90% of the patients, and rerun the biclustering algorithm, and obtain 20 new biclusterings. Then, we calculate the stability of each density-invariant bicluster found in the previous step using average best-match score. First, we calculate the match score of two biclusters $B_1 = (G_1, C_1)$, $B_2 = (G_2, C_2)$, where G_1, G_2 represent gene sets and C_1, C_2 represent condition sets. Similar to Prelic et al. and Lie et al. [19,28], the match score of biclusters $B_1 = (G_1, C_1)$, $B_2 = (G_2, C_2)$ is calculated as follows:

$$\texttt{match}(B_1, B_2) = \frac{|G_1 \cap G_2| + |C_1 \cap C_2|}{|G_1 \cup G_2| + |C_1 \cup C_2|}\,. \tag{19}$$

Let $M = \bigcup_{i=1}^{k} B_i^*$ be a biclustering of the subsample of the dataset. We compare a bicluster $B = (G, C)$ to all biclusters in $B_i^* \in M$, and assign the maximum match value as the best-match score:

$$\texttt{best\_match}(B, M = \bigcup_{i=1}^{k} B_i^*) = \max_{B_i^* \in M} \texttt{match}(B, B_i^*)\,. \tag{20}$$

Finally, we take the average best-match score of each bicluster B by comparing them to each biclustering M_i, and obtain the average best-match score of bicluster B as follows:

# Configuration	Method	DIB parameters (th, α, β)	# DIB
1	MBF	0.80, 0.80, 0.15	8
2	TBF (PARAFAC)	0.98, 0.89, 0.01	170
	TBF (Tucker3)	0.60, 0.60, 0.40	5
3	CMTBF$_g$ (CMTF_PARAFAC_ALS)	0.60, 0.60, 0.40	0
	CMTBF$_g$ (CMTF_Tucker_ALS)	0.70, 0.70, 0.30	4
4	CMTBF$_s$ (CMTF_PARAFAC_ALS)	0.80, 0.90, 0.10	21
	CMTBF$_s$ (CMTF_Tucker_ALS)	0.80, 0.85, 0.15	6
5	CMTBF$_{gs}$ (CMTF_PARAFAC_ALS)	0.98, 0.99, 0.01	0
	CMTBF$_{gs}$ (CMTF_Tucker_ALS)	0.60, 0.70,0.30	5
6	CMMBF	0.60, 0.60, 0.40	17

Table 1: Biclustering results for each data configuration, including density-invariant biclustering algorithm parameters and number of density-invariant biclusters (DIB). For TBF, PARAFAC and Tucker3 model, results are listed separately. Similarly, for CMTBF, CMTF_PARAFAC_ALS and CMTF_Tucker_ALS, results are listed separately. When there are no stable biclusters with average best-match score $\geq 95\%$, five most stable biclusters are picked as the stable biclusters.

$$\texttt{average\_best\_match}(B, \bigcup_{i=1}^{n} M_i) = \frac{\sum_{i=1}^{n} \texttt{best\_match}(B, M_i)}{n}. \tag{21}$$

We pick the biclusters with $\geq 95\%$ average best-match score as statistically significant biclusters, and evaluate their biological relevance. If there are no significant biclusters, we report top 5 stable biclusters with their average best-match scores.

4 Results

In order to find host-pathogen associations in tuberculosis patient dataset, we biclustered spoligotypes and countries using six different data configurations shown in Figure 3. For each data configuration, we followed the steps of Unified Biclustering Framework (UBF), and found the most stable biclusters. Table 1 shows the parameters of `DensityInvariantBiclustering` (th, α, β) and number of density-invariant biclusters for each data configuration. Note that PARAFAC and Tucker3 variants of TBF, CMTF_PARAFAC_ALS and CMTF_Tucker_ALS variants of CMTBF are listed separately. Next, we evaluate the statistical significance and biological relevance of biclusters for each data configuration, and find host-pathogen associations within the whole patient dataset and within each major lineage.

4.1 Biclusters using spoligotypes and country of birth

We first contract and sum the host-pathogen tensor along the time mode and find biclusters based on the distribution of spoligotypes to countries of birth, as in data configuration 1 in Figure 3. In this setting, no distance measure or time is added to the domain knowledge. Table 2 shows the density-invariant biclusters. Bicluster B1 suggests that patients from Haiti are infected with ST1162 strain, a Beijing strain, and ST398, a LAM4 strain. Bicluster B12 is listed in the supplementary material due to its size: http://tbinsight.cs.rpi.edu/UBFsupp.rar. This bicluster contains 848 patients from United States who are infected with 63 different strains. One of these strains is the transmissive Beijing strain ST1 which initiated many outbreaks

Bicluster	Number of patients	Spoligotypes			Countries	
		SIT no	Major lineage	Sublineage	Name	TB continent
B11	5	ST1162 ST398	East-Asian Euro-American	Beijing LAM4	Haiti	Americas
B13	19	ST265 ST422 ST89 ST287 ST1268 ST25 ST732	East-Asian *M. bovis* Indo-Oceanic Indo-Oceanic Euro-American East-African Indian Euro-American	Beijing BOV_1 EAI5 EAI2-Manila T5 CAS1-Delhi T1	China	East Asia
B14	6	ST1908 ST58	Euro-American Euro-American	H3 T5	Ecuador	Americas
B15	6	ST43 ST848 ST511	Indo-Oceanic Euro-American Euro-American	EAI6-BGD1 T2 H3	Dominican Republic	Americas
B16	2	ST897	Indo-Oceanic	EAI2-Manila	Philippines	Southeast Asia
B17	2	ST447	Euro-American	T1	Bangladesh	Indian Subcontinent
B18	4	UST251 ST1154	Euro-American Euro-American	S LAM9	Mexico	Americas

Table 2: Biclustering results on data configuration 1 using UBF. Biclusters associate spoligotypes to country of birth of patients. For spoligotypes, SIT number, major lineage based on CBN, and sublineage based on KBBN are listed. For countries, the name and the TB continent are listed. Bicluster B16 represents the well-known association between patients from Philippines and EAI2-Manila strains.

in United States [29, 30]. Bicluster B13 shows that patients from China are infected with 7 different strains. Bicluster B14 shows that ST1908 and ST58 are two Ecuadorian isolates belonging to Euro-American lineage. Bicluster B16 is a well-known association, and suggests that patients from Philippines are infected with an EAI2-Manila strain, ST897. Bicluster B18 suggests that Mexican patients, as neighbours of United States, are infected with UST251 and ST1154, two Euro-American strains. The five most stable biclusters are B16, B17, B18, B11, B14, and their average best-match scores are in the range [0.1667, 0.2]. One may argue that biclusters with few patients does not constitute a strong host-pathogen association. This suggests that TB detection rate should be increased to gather more patient data and make more accurate inferences on host-pathogen association.

4.2 Incorporating time

The original host-pathogen tensor has time as the third mode. Therefore, when we found biclusters using the host-pathogen tensor as in data configuration 2 of Figure 3, we account for distribution of spoligotypes to countries of birth through time, in this case years from 2001 to 2007. When we use PARAFAC to decompose the host-pathogen tensor, we found 170 density-invariant biclusters. Here, we focus on five most stable biclusters when PARAFAC model is used. Average best-match scores of these five biclusters range from 0.6915 to 0.7295. The full list of these biclusters can be found in the supplementary material. Bicluster B211 associates Vietnamese patients to 11 strains belonging to Euro-American, East Asian, Indo-Oceanic and East-African Indian lineages. Bicluster B212 suggests that patients from Peru are infected with 17 different strains belonging to Euro-American, Indo-Oceanic, and East Asian lineages. Bicluster B214 is shown in Table 3. There are 111 patients in bicluster B214 from India, Peru and Vietnam, which are infected with 6 Euro-American strains and one East Asian strain. Notice that this East Asian strain is ST1, which is the transmissive Beijing strain. This suggests that some of the patients in this bicluster must be involved in the outbreaks in United States initiated by ST1 Beijing strains.

When Tucker3 model is used to decompose the host-pathogen tensor, we find 5 density-invariant biclusters. Their average-best match scores range from 0.04 to 0.18, which shows that biclusters found using Tucker3 model are less stable compared to the ones found using PARAFAC model. These five biclusters, bicluster B221 to B225, are listed in the supplementary material. Bicluster B221 suggests that US patients are

Bicluster	Number of patients	Spoligotypes			Countries	
		SIT no	Major lineage	Sublineage	Name	TB continent
B214	111	ST53 ST17 ST1 ST197 ST61 ST119 ST42	Euro-American Euro-American East Asian Euro-American Euro-American Euro-American Euro-American	T1 LAM2 Beijing X3 LAM10-CAM X1 LAM9	India Peru Vietnam	Indian Subcontinent Americas Southeast Asia
B225	4	ST294 ST290 ST176	Euro-American Euro-American Euro-American	H3 LAM9 LAM6	Mexico	Americas

Table 3: Biclustering results on data configuration 2 using PARAFAC and Tucker3 models via UBF on the host-pathogen tensor. Bicluster B214 associates patients from India, Peru and Vietnam to 6 Euro-American strains and the transmissive East Asian Beijing strain ST1. Bicluster B224 groups Mexican patients infected with three different Euro-American strains.

infected with 31 different strains, and bicluster B222 suggests that Chinese patients are infected with 11 strains belonging to Euro-American, East-African Indian, and *M. bovis* lineages. Note that no East Asian strain is associated with Chinese patients, which suggests that our biclustering analysis on the host-pathogen tensor is introducing noise when time is added into domain knowledge. Bicluster B225, also listed in Table 3, has 4 patients and suggests that Mexican patients are infected with ST294, ST290 and ST176 strains, all members of Euro-American lineage.

4.3 Incorporating time and distance measures

Next, we concatenate distance matrices one at a time, and finally both of them, to the host-pathogen tensor. Concatenation of genetic proximity matrix results in data configuration 3, concatenation of spatial proximity matrix results in data configuration 4, and concatenation of both matrices results in data configuration 5. We factorize these matrices using coupled matrix-tensor factorization via CMTF_PARAFAC_ALS and CMTF_Tucker_ALS, and report statistically significant and biologically relevant biclusters. The full list of biclusters can be found in the supplementary material.

If we use genetic distance matrix only, factorization via CMTF_PARAFAC_ALS results in no density-invariant biclusters. When the coupled matrix-tensor is decomposed via CMTF_Tucker_ALS, 4 stable clusters are found, the stability of which range from 0.08 to 0.19. Two of these biclusters, B321 and B323, are listed in Table 4. Bicluster B321 groups 32 patients from Ecuador, infected with ST53, ST62, ST51, ST1908, which are all Euro-American strains. Notice that ST53 and ST51 belong to T1 sublineage, which is a class of ill-defined Euro-American strains. Bicluster B323 contains 4 patients from Mexico, all infected with ST52, a Euro-American T2 strain.

If we use spatial distance matrix only, factorization via CMTF_PARAFAC_ALS results in 21 density-invariant biclusters, and we picked 5 most stable biclusters among them. The stability values of these biclusters range from 0.26 to 0.32. Table 4 shows 3 of these biclusters. Bicluster B411 suggests that Euro-American LAM2 strain ST908 infects patients from Dominican Republic, Puerto Rico, Trinidad Tobago, and Unites States, all from Americas. Notice how geographically close countries are collected together in a bicluster in the host-pathogen association analysis by incorporating spatial proximity into domain knowledge. Bicluster B412 suggests that Euro-American T5 strain ST904 infects patients from Ecuador, Haiti, Trinidad Tobago, and United States, which are again all in Americas. Bicluster B414 includes strains of both Bicluster B411 and B412, and combines the two common countries in these biclusters. It suggests that Euro-American T5 strain ST904 and Euro-American LAM2 strain ST908 infect patients from Trinidad Tobago and United States. When we factorize the coupled matrix-tensor via CMTF_Tucker_ALS in UBF, 6 density-invariant biclusters

Bicluster	Number of patients	Spoligotypes			Countries	
		SIT no	Major lineage	Sublineage	Name	TB continent
B321	32	ST53	Euro-American	T1	Ecuador	Americas
		ST62	Euro-American	H1		
		ST51	Euro-American	T1		
		ST1908	Euro-American	H3		
B323	4	ST52	Euro-American	T1	Mexico	Americas
B411	6	ST908	Euro-American	LAM2	Dominican Rep.	Americas
					Puerto Rico	Americas
					Trinidad and Tobago	Americas
					United States	Americas
B412	6	ST904	Euro-American	T5	Ecuador	Americas
					Haiti	Americas
					Trinidad and Tobago	Americas
					United States	Americas
B414	6	ST904	Euro-American	T5	Trinidad and Tobago	Americas
		ST908	Euro-American	LAM2	United States	Americas
B421	32	ST1	East Asian	Beijing	Taiwan	East Asia
					Barbados	Americas
					Dominica	Americas
					Malaysia	Southeast Asia
					Myanmar	Southeast Asia
					Philippines	Southeast Asia
B422	27	ST1	East Asian	Beijing	Malaysia	Southeast Asia
		ST38	Euro-American	X2	Philippines	Americas
B425	2	ST93	Euro-American	LAM5	Honduras	Americas
B525	11	ST167	Euro-American	T1	Haiti	Americas
		ST42	Euro-American	LAM9		
		ST57	Euro-American	LAM10-CAM		
		ST904	Euro-American	T5		
		ST187	*M. africanum*	AFRI_1		
		ST1867	*M. africanum*	AFRI_1		

Table 4: Biclustering results on data configuration 3, 4, 5 using CMTF_PARAFAC_ALS and CMTF_Tucker_ALS algorithms via UBF on the coupled matrix-tensor. Biclusters B411 and B412 suggests that Euro-American strains ST908 and ST904 infects patients from four spatially close countries in Americas respectively. Bicluster B421 suggests that transmissive Beijing strain ST1 is wide-spread and infects patients from three different TB continents. Bicluster B422 groups patients from two neighbour countries, Malaysia and Philippines, who are infected with Beijing strain ST1 and X2 strain ST38.

are found, and we picked 5 most stable biclusters among them, with average best-match score ranging from 0.09 to 0.50. Table 4 shows 3 of these biclusters. Bicluster B421 points out that transmissive ST1 Beijing strain is wide-spread, and it infects patients from Taiwan, Barbados, Dominica, Malaysia, Myanmar, and Philippines, which cover 3 different TB continents: East Asia, Americas, and Southeast Asia. This shows that, even if we use spatial proximity matrix to narrow down transmission events, transmissive ST1 strain is still associated with patients from multiple TB continents. Bicluster B422 contains 27 patients from Philippines and Malaysia, both from Southeast Asia, which are infected with ST1 and ST38 strains. Notice how these countries are grouped together using the spatial proximity matrix. Bicluster B425 consists of 2 patients from Honduras, both infected with Euro-American LAM5 strain ST93.

If we concatenate both genetic and spatial proximity matrices, factorization via CMTF_PARAFAC_ALS does not assign any density-invariant biclusters. When the coupled matrix-tensor is decomposed via CMTF_Tucker_ALS, we find 5 density-invariant biclusters, with average best-match score values ranging from 0.11 to 0.30. Table 4 shows one of these biclusters. Bicluster B525 contains 11 patients from Haiti which are infected with Euro-American strains ST167, ST42, ST57, ST904, and *M. africanum* AFRI_1 strains ST187 and ST1867. The full list of biclusters can be found in supplementary material. Notice that there is no order in stability of biclusters found using CMTF_PARAFAC_ALS and CMTF_Tucker_ALS. However, biclusters found using CMTF_Tucker_ALS are more biologically coherent. This shows that that high stability does not imply biological relevance.

Bicluster	Number of patients	Spoligotypes			Countries	
		SIT no	Major lineage	Sublineage	Name	TB continent
B64	3	ST1391	Indo-Oceanic	EAI5	Bangladesh	Indian Subcontinent
		ST58	Euro-American	T1		
B66	19	ST1162	East Asian	Beijing	Haiti	Americas
		ST168	Euro-American	H3		
		ST398	Euro-American	LAM4		
		ST57	Euro-American	LAM10-CAM		
		ST874	Euro-American	S		
		UST256	Euro-American	H1		
		ST541	East Asian	Beijing		
		ST1867	*M. africanum*	AFRI_1		
		ST822	Euro-American	LAM9		
		ST546	Euro-American	X3		
		ST3	Euro-American	LAM2		

Table 5: Biclustering results on data configuration 6 using CMMF_ALS via UBF on the coupled matrix-matrix. Bicluster B64 groups patients from Bangladesh who are infected with two strains of ill-defined sublineages: Indo-Oceanic EAI5 strain ST1391 and Euro-American T1 strain ST58.

4.4 Incorporating distance, but not time

Finally, in the last data configuration, we use genetic distance, spatial distance, but not time. This reduces the mutation path length and transmission path length, which increases the likelihood of mutation between the set of strains and transmission between the set of patients. To do so, we contract and sum the host-pathogen tensor along the time mode, and concatenate the genetic proximity matrix and spatial proximity matrix. We bicluster spoligotypes and countries using CMMF_ALS on this dataset in UBF. There are 17 density-invariant biclusters, and we picked the ones with average best-match score of 90% and above. Full list of these biclusters are in the supplementary material. Table 5 shows two of these biclusters, B64 and B66. Bicluster B64 contains 3 patients from Bangladesh infected with Indo-Oceanic EAI5 strain ST1391 and Euro-American T1 strain ST447. Notice that EAI5 is a generic sublineage of Indo-Oceanic lineage, and T1 is a generic sublineage of Euro-American lineage, and they are both ill-defined. Bicluster B66 contains 19 patients from Haiti infected with Euro-American, East Asian and *M. africanum* strains. Haiti is an island next to Dominican Republic and immigrants of Haiti must have brought strains belonging to various lineages.

4.5 Host-pathogen association within each major lineage

The six phylogeographic major lineages determined by CBN are established. Therefore, we subdivide the patient dataset based on six major lineages, and run UBF on each of them. We used data configuration 6, since it resulted in both stable and biologically relevant biclusters in the complete patient dataset. We found the most stable host-pathogen associations for each major lineage and reported their biological relevance.

Table 6 shows some of the most stable and biologically relevant biclusters. The full list of biclusters can be found in the supplementary material. Bicluster B711 of Euro-American lineage in the list of supplementary material contains 628 US patients infected with 61 different strains. Bicluster B712 listed in Table 6 suggests a strong association between Mexican patients and pathogens of three Euro-American strains, S strain UST251, X2 strain ST478, and LAM9 strain ST1154. Notice that all strains belong to different sublineages of Euro-American lineage. The average best-match score of this bicluster is 0.7783. Bicluster B721 of Indo-Oceanic lineage listed in the supplementary material suggests an association between 40 Chinese patients and 16 different Indo-Oceanic strains, belonging to various sublineages. The stability value of 0.9621 suggests that this is a strong host-pathogen association.

Bicluster B732 listed in Table 6 contains 9 Chinese patients infected with CAS1-Delhi, CAS and EAI5 strains of East-African Indian lineage. Similarly, bicluster B733 suggests that patients from China and Dominican Republic are likely to be infected with the following East-African Indian strains: CAS1-Delhi strains ST381

Bicluster	Number of patients	Spoligotypes			Countries	
		SIT no	Major lineage	Sublineage	Name	TB continent
B712	5	UST251	Euro-American	S	Mexico	Americas
		ST478	Euro-American	X2		
		ST1154	Euro-American	LAM9		
B732	9	ST471	East-African Indian	CAS1-Delhi	China	East Asia
		ST25	East-African Indian	CAS1-Delhi		
		ST381	East-African Indian	CAS1-Delhi		
		ST21	East-African Indian	CAS		
		ST203	East-African Indian	CAS		
		UST167	East-African Indian	EAI5		
B733	11	ST381	East-African Indian	CAS1-Delhi	China Dominican Republic	East Asia Americas
		ST25	East-African Indian	CAS1-Delhi		
		ST21	East-African Indian	CAS		
		UST167	East-African Indian	EAI5		
B741	7	ST1162	East Asian	Beijing	Haiti	Americas
		ST941	East Asian	Beijing		
		ST541	East Asian	Beijing		
		ST1168	East Asian	Beijing		
B742	212	UST1	East Asian	Beijing	United States	Americas
		ST255	East Asian	Beijing		
		ST260	East Asian	Beijing		
		ST941	East Asian	Beijing		
		ST265	East Asian	Beijing		
		ST190	East Asian	Beijing		
		ST1	East Asian	Beijing		
B743	291	ST260	East Asian	Beijing	China United States	East Asia Americas
		ST265	East Asian	Beijing		
		ST1	East Asian	Beijing		
B751	17	ST325	*M. africanum*	AFRI_1	United States	Americas
		ST326	*M. africanum*	AFRI_1		
		ST187	*M. africanum*	AFRI_1		
		ST181	*M. africanum*	AFRI_1		
		ST319	*M. africanum*	AFRI_2		
		ST331	*M. africanum*	AFRI_2		
		UST229	*M. africanum*	AFRI_2		
B761	3	ST479	*M. bovis*	BOV	Dominican Republic	Americas
		ST481	*M. bovis*	BOV_1		
B762	9	ST409	*M. bovis*	BOV_2	United States	Americas
		ST683	*M. bovis*	BOV_2		

Table 6: Biclustering results on data configuration 6 using CMMF_ALS via UBF on the coupled matrix-matrix for each major lineage. Bicluster B712 suggests that Mexican patients are likely to be infected with UST251, ST478, and ST1154 strains, given that the pathogen is a Euro-American strain. Bicluster B742 groups 212 US patients and shows that US patients are commonly infected with Beijing strains, including the transmissive ST1 strain. 291 patients in bicluster B743 shows that Beijing strains ST260, ST265 and the transmissive ST1 strain infects both Chinese and US patients. Biclusters B761 and B762 suggest that, given that MTBC is an *M. bovis* strain, it is more likely to infect a patient from Dominican Republic if it is a BOV or BOV_1 strain, and more likely to infect a US patient if it is a BOV_2 strain.

and ST25, CAS strain ST21, and EAI5 strain UST167. Bicluster B741 suggests that Haitian patients are infected with the following Beijing strains: ST1162, ST941, ST541, ST1168. Similarly, 212 US patients in bicluster B742 suggests that US patients are infected commonly with the following Beijing strains: UST1, ST255, ST260, ST941, ST265, ST190, and the transmissive ST1 strain. 291 patients in bicluster B743 suggest that both Chinese and US patients are infected with the following Beijing strains very frequently: ST260, ST265, and the transmissive ST1 strain. This shows that Beijing strains brought to the US by Chinese immigrants infect both Chinese and US patients in the US.

Bicluster B751 shows that US patients are infected with AFRI_1 strains ST325, ST326, ST187, ST181 and AFRI_2 strains ST319, ST331, UST229 of *M. africanum* lineage. Bicluster B761 suggests that patients from Dominican Republic are likely to be infected with BOV strain ST479 and BOV_1 strain ST481 belonging to *M. bovis* lineage. On the other hand, bicluster B762 suggests that US patients are infected BOV_2 strains ST409 and ST683 belonging to *M. bovis* lineage. These two biclusters suggest that, given an *M. bovis* strain, it is likely to infect a patient from Dominican Republic if it is a BOV or BOV_1 strain, whereas it is more likely to have infected a US patient if the strain is a BOV_2 strain.

5 Discussion and Conclusion

We developed the Unified Biclustering Framework (UBF) to find host-pathogen associations in tuberculosis patients. To our knowledge, this is the first study to restate host-pathogen association analysis as a biclustering problem. UBF is flexible in the sense that distance and time can be added into domain knowledge of data analysis via coupled matrix-matrix and matrix-tensor factorization. This enables genome-phenome data fusion in one unsupervised learning framework.

Each bicluster refers to a possible host-pathogen association. We found statistically significant biclusters, some of which represent well-known host-pathogen relationships and some of which reveal new associations. For instance, bicluster B16 shows the well-known association of patients from Philippines and EAI2-Manila strains. Similarly, biclusters B742 and B743 shows that many US patients are infected with Beijing strains including ST1 strain, a well-known initiator of many outbreaks in the US. On the other hand, we also found new patient-strain relationships via genome-phenome data fusion by adding genetic proximity, spatial proximity and time into domain knowledge. For instance, bicluster B422 groups patients from two neighbour countries, Malaysia and Philippines, who are infected with Beijing strain ST1 and X2 strain ST38. Biclusters B761 and B762 suggest that patients from Dominican Republic are infected with BOV and BOV_1 strains of *M. bovis* lineage, whereas US patients are infected with BOV_2 strains of *M. bovis* lineage. Note that although we picked statistically significant biclusters, statistical significance does not imply biological relevance [31]. However, these new stable biclusters lead to new host-pathogen associations.

Host-pathogen association analysis can be extended by adding new patient and strain attributes. As future work, we will add MIRU and RFLP, two biomarkers of MTBC, into this analysis. In addition, we will add other patient attributes such as age group, ethnicity, homelessness and other risk factors of TB. We will also speed up UBF using line search in ALS-based coupled factorization algorithms. This will enhance both the speed and accuracy of coupled factorizations, which will lead to more accurate host-pathogen associations.

Acknowledgments

We thank Dr. Minoo Aminian for providing the KBBN sublineage labels of strains. This work was supported by NIH R01LM009731.

References

[1] World Health Organization (WHO) Report, "*Global tuberculosis control : epidemiology, strategy, financing*," 2009.

[2] A. E. Hirsh, A. G. Tsolaki, K. DeRiemer, M. W. Feldman, and P. M. Small, "Stable association between strains of *Mycobacterium tuberculosis* and their human host populations," *Proc. Nat. Academy of Sciences (PNAS)*, vol. 101, no. 14, pp. 4871–4876, 2004.

[3] S. Gagneux, K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, M. Hilty, P. C. Hopewell, and P. M. Small, "Variable host-pathogen compatibility in *Mycobacterium tuberculosis*," *Proc. Nat. Academy of Sciences (PNAS)*, vol. 103, no. 8, pp. 2869–2873, 2006.

[4] K. Bennett, C. Ozcaglar, J. Ranganathan, S. Raghavan, J. Katz, D. Croft, B. Yener, and A. Shabbeer, "Visualization of tuberculosis patient and *Mycobacterium tuberculosis* complex genotype data via host-pathogen maps," in *Proc. 2011 IEEE Int. Conf. Bioinformatics and Biomedicine Workshops (BIBMW)*, pp. 124–129, nov. 2011.

[5] S. Yu, T. Falck, A. Daemen, L.-C. Tranchevent, J. Suykens, B. De Moor, and Y. Moreau, "L2-norm multiple kernel learning and its application to biomedical data fusion," *BMC Bioinformatics*, vol. 11, no. 1, p. 309, 2010.

[6] G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, 2004.

[7] G. R. G. Lanckriet, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based integration of genomic data using semidefinite programming," in *Kernel Methods in Computational Biology*, pp. 231–263, Cambridge, MA: MIT Press, 2004.

[8] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[9] C. Ozcaglar, A. Shabbeer, S. L. Vandenberg, B. Yener, and K. P. Bennett, "Sublineage structure analysis of *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors," *BMC Genomics*, vol. 12, no. Suppl 2, p. S1, 2011.

[10] K. Lage, E. O. Karlberg, Z. M. Storling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[11] 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. Microbiology*, vol. 35, no. 4, pp. 907–914, 1997.

[12] C. Ozcaglar, A. Shabbeer, N. Kurepina, B. Yener, and K. P. Bennett, "Data-driven insights into deletions of *Mycobacterium tuberculosis* complex chromosomal DR Region using spoligoforests," in *Proc. 2011 IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, (Atlanta, GA), pp. 75–82, 2011.

[13] C. Ozcaglar, A. Shabbeer, N. Kurepina, N. Rastogi, B. Yener, and K. P. Bennett, "Inferred spoligoforest topology unravels spatially bimodal distribution of mutations in the DR region." *IEEE Trans. NanoBioscience*, in press, 2012.

[14] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proc. Intelligent Systems for Molecular Biology (ISMB)*, pp. 93–103, 2000.

[15] I. S. Dhillon, “Co-clustering documents and words using bipartite spectral graph partitioning,” in *Proc. seventh ACM SIGKDD Int. Conf. Knowledge discovery and data mining (KDD ’01)*, pp. 269–274, 2001.

[16] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, “Spectral biclustering of microarray data: coclustering genes and conditions,” *Genome Research*, vol. 13, no. 4, pp. 703–716, 2003.

[17] A. Tanay, R. Sharan, and R. Shamir, “Discovering statistically significant biclusters in gene expression data,” *Bioinformatics*, vol. 18, no. suppl 1, pp. S136–S144, 2002.

[18] G. Getz, E. Levine, and E. Domany, “Coupled two-way clustering analysis of gene microarray data,” *Proc. Nat. Academy of Sciences (PNAS)*, vol. 97, no. 22, pp. 12079–12084, 2000.

[19] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, “A systematic comparison and evaluation of biclustering methods for gene expression data,” *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[20] R. Colak, F. Moser, J. S.-C. Chu, A. Schonhuth, N. Chen, and M. Ester, “Module discovery by exhaustive search for densely connected, co-expressed regions in biomolecular interaction networks,” *PLoS ONE*, vol. 5, p. e13348, Oct. 2010.

[21] S. C. Madeira and A. L. Oliveira, “Biclustering algorithms for biological data analysis: A survey,” *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 1, pp. 24–45, Jan. 2004.

[22] 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.

[23] M. Aminian, A. Shabbeer, K. Hadley, C. Ozcaglar, S. L. Vandenberg, and K. P. Bennett, “Knowledge-based Bayesian network for the classification of *Mycobacterium tuberculosis* complex sublineages,” in *Proc. 2nd ACM Conf. Bioinformatics, Computational Biology and Biomedicine*, pp. 201–208, 2011.

[24] C. Demay, B. Liens, T. Burguiére, 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,” *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 755 – 766, 2012.

[25] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms.* Cambridge, MA: The MIT Press, 2001.

[26] E. Acar, T. G. Kolda, and D. M. Dunlavy, “All-at-once optimization for coupled matrix and tensor factorizations,” *ArXiv e-prints*, May 2011.

[27] P. Dao, R. Colak, R. Salari, F. Moser, E. Davicioni, A. Schonhuth, and M. Ester, “Inferring cancer subnetwork markers using density-constrained biclustering,” *Bioinformatics*, vol. 26, no. 18, pp. i625–i631, 2010.

[28] X. Liu and L. Wang, “Computing the maximum similarity biclusters of gene expression data,” *Bioinformatics*, vol. 23, no. 1, pp. 50–56, 2007.

[29] P. J. Bifani, B. Mathema, N. E. Kurepina, and B. N. Kreiswirth, “Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains,” *Trends in Microbiology*, vol. 10, no. 1, pp. 45 – 52, 2002.

[30] J. R. Glynn, J. Whiteley, P. J. Bifani, K. Kremer, and D. van Soolingen, “Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review,” *Emerging Infectious Diseases*, vol. 8, no. 8, pp. 843 – 849, 2002.

[31] M. Zervakis, M. Blazadonakis, G. Tsiliki, V. Danilatou, M. Tsiknakis, and D. Kafetzopoulos, “Outcome prediction based on microarray analysis: a critical perspective on methods,” *BMC Bioinformatics*, vol. 10, no. 1, p. 53, 2009.