Data-driven insights into deletions of *Mycobacterium tuberculosis* complex chromosomal DR region using spoligoforests

Cagri Ozcaglar[1], Amina Shabbeer[1], Natalia Kurepina[3], Bülent Yener[1], Kristin P. Bennett[1,2]
(1) Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY
(2) Mathematical Sciences Department, Rensselaer Polytechnic Institute, Troy, NY
(3) TB Center, Public Health Research Institute, UMDNJ, Newark, NJ
ozcagc2@cs.rpi.edu, shabba@cs.rpi.edu, kurepina@umdnj.edu, yener@cs.rpi.edu, bennek@rpi.edu

***Abstract*—Biomarkers of *Mycobacterium tuberculosis* complex (MTBC) mutate over time. Among the biomarkers of MTBC, spacer oligonucleotide type (spoligotype) and Mycobacterium Interspersed Repetitive Unit (MIRU) patterns are commonly used to genotype clinical MTBC strains. In this study, we present an evolution model of spoligotype rearrangements using MIRU patterns to disambiguate the ancestors of spoligotypes, in a large patient dataset from the United States Centers for Disease Control and Prevention (CDC). Based on the contiguous deletion assumption and rare observation of convergent evolution, we first generate the most parsimonious forest of spoligotypes, called a spoligoforest, using three genetic distance measures. An analysis of topological attributes of the spoligoforest and number of variations at the direct repeat (DR) locus of each strain reveals interesting properties of deletions in the DR region. First, we compare our mutation model to existing mutation models of spoligotypes and find that our mutation model produces as many within-lineage mutation events as other models, with slightly higher segregation accuracy. Second, based on our mutation model, the number of descendant spoligotypes follows a power law distribution. Third, contrary to prior studies, the power law distribution does not plausibly fit to the mutation length frequency. Finally, the total number of mutation events at consecutive DR loci follows a bimodal distribution, which results in accumulation of shorter deletions in the DR region. The two modes are spacers 13 and 40, which are hotspots for chromosomal rearrangements. The change point in the bimodal distribution is spacer 34, which is absent in most MTBC strains. This bimodal separation results in accumulation of shorter deletions, which explains why a power law distribution is not a plausible fit to the mutation length frequency.**

***Keywords*-tuberculosis, *Mycobacterium tuberculosis* complex, DR locus, spoligotype, MIRU-VNTR, mutation.**

I. Introduction

Tuberculosis (TB) is a leading cause of death among infectious diseases. Tuberculosis is caused by *Mycobacterium tuberculosis* complex (MTBC). One third of the human population is infected, either latently or actively, with MTBC bacteria [1]. DNA fingerprinting of MTBC strains is used for tracking and understanding the transmission of tuberculosis. Isolates from TB patients are genotyped using multiple biomarkers, which include spacer oligonucleotide types (spoligotypes), Mycobacterium Interspersed Repetitive Units - Variable Number Tandem Repeats (MIRU-VNTR), and IS*6110* Restriction Fragment Length Polymorphism (RFLP) [2], [3], [4].

Biomarkers of MTBC change over time. Brosch et al. presented an evolutionary repetition model based on the analysis of twenty regions of difference (RD) found in a comparison of whole genome sequences of MTBC clinical strains [5], [6]. Tanaka et al. introduced cluster-graphs to analyze genotype clusters of MTBC separated by a single mutation step [7]. Based on the observation that deletion length follows a Zipf distribution, Reyes et al. presented a probabilistic mutation model of spoligotypes to disambiguate the ancestors [8]. Grant et al. simulated stepwise loss or gain of repeats in MIRU loci using a stochastic continuous-time model, and suggested that all MIRU loci mutate very slowly [9].

In this study, we present a mutation model of spoligotypes based on variations in the direct repeat (DR) region. To disambiguate the parents in the cluster-graph, we add an independent biomarker, MIRU-VNTR. First, we use a large patient dataset from the United States Centers for Disease Control and Prevention (CDC) and generate the most parsimonious forest of spoligotypes, called a spoligoforest. The spoligoforest generation is based on the contiguous deletion assumption, nonexistence of convergent evolution and three distance measures defined on spoligotypes and MIRU patterns. The spoligoforest of the CDC dataset in Figure 1 generated using this model contains the putative history of mutation events in the chromosomal DR region. Each node in the spoligoforest represents a distinct spoligotype, and each edge represents a potential mutation event from parent spoligotype to child spoligotype. The number of spacers lost in a mutation event is referred as the mutation length. We compare the DR evolution model to existing mutation models in terms of number of mutations and segregation accuracy and show that our mutation model with the additional biomarker, MIRU-VNTR, leads to as many within-lineage mutation events as in other mutation models. We identified topological attributes of the spoligoforest and gave insights into variations of spoligotypes. Based on the spoligoforest, the number of descendant spoligotypes follows a power law distribution. On the other hand, based on goodness-of-fit results, mutation length frequency does not follow a power law distribution. The number of mutations at contiguous DR loci follows a bimodal distribution, and the modes are spacer 13 and spacer 40, which are hotspots, e.g. sites of increased observed variability. Spacer 34 is the change point in the distribution, and it is stable, which is due to lack of spacers 33-36 in most MTBC strains in the CDC dataset. We hypothesize that this bimodal distribution results in unobservable longer mutation events, which is why power law distribution is not a plausible fit to mutation length frequency.

II. Background

In order to build a mutation model for evolution of the chromosomal DR region, we used two biomarkers of MTBC: spoligotypes and MIRU patterns. Each biomarker has a different mutation mechanism which is analyzed separately. Spoligotypes can lose spacers in the DR region, but not gain, while MIRU loci can either lose or gain tandem repeats [10], [11], [12]. In this section, we

978-0-7695-4574-5/11 $26.00 © 2011 IEEE
DOI 10.1109/BIBM.2011.64

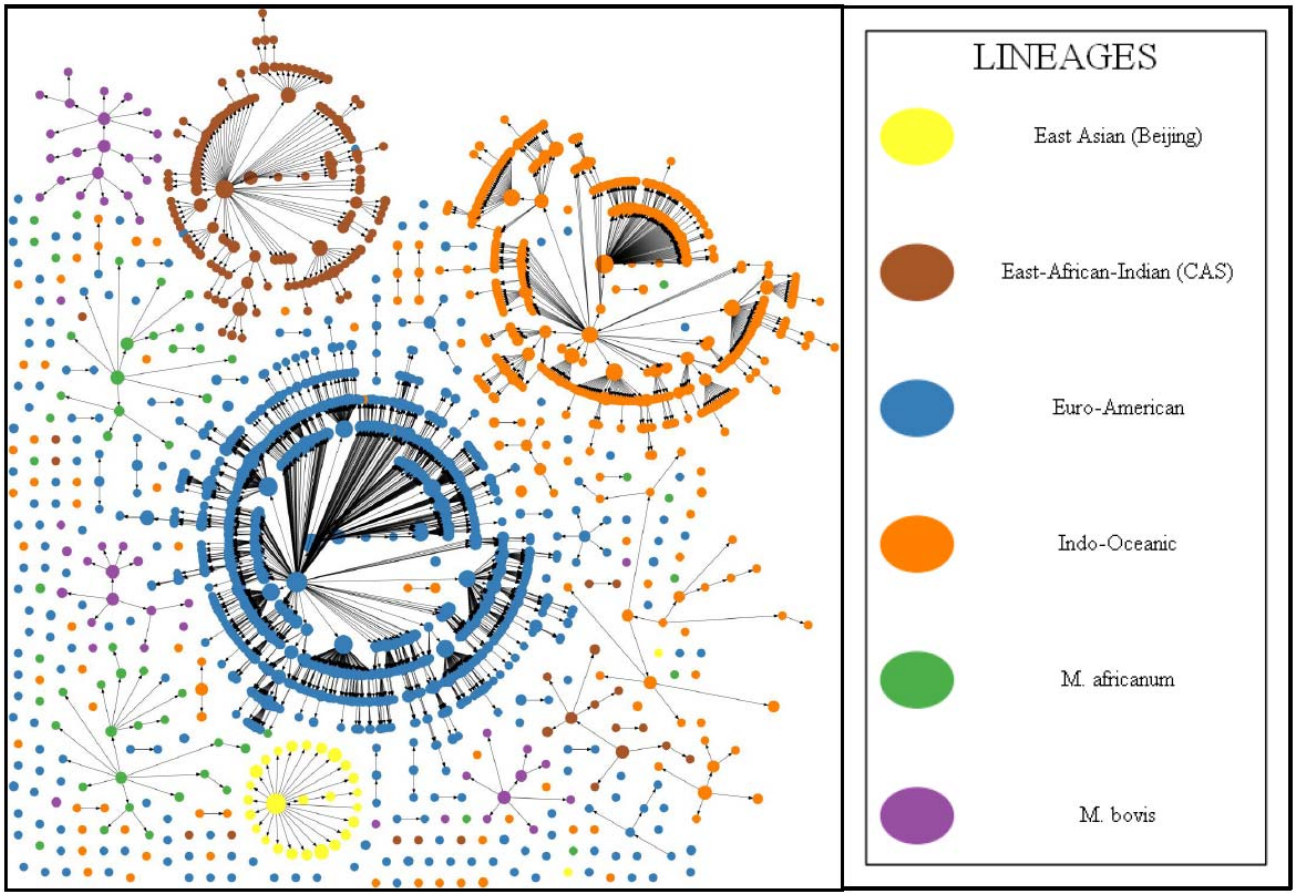

Figure 1: The spoligoforest of the CDC dataset. Each node represents a distinct spoligotype, and each edge represents a one-step mutation event from parent spoligotype to child spoligotype. Node sizes are proportional to the number of patients infected with MTBC strains having the spoligotype, in log scale. Nodes are colored by major lineages of MTBC strains. The spoligoforest generator is implemented in Java, using the visualization software Graphviz [13].

give a brief background on spoligotyping, MIRU-VNTR typing, and mutation of both biomarkers.

A. *Spoligotyping*

Spoligotyping is a PCR-based genotyping method of MTBC that exploits the polymorphism in the DR locus, which is a member of the clustered regularly interspaced palindromic repeats (CRISPR) loci [14]. The DR region is a polymorphic locus which comprises of directly repeating sequences of 36 bp, separated by unique spacer sequences of 36 to 41 bp [15]. One repeat sequence and the following spacer sequence together is termed a direct variable repeat (DVR). A spoligotype is composed of 43 spacers, which are represented by a 43-bit binary sequence, where zeros and ones indicate absence and presence of particular spacer in the DR locus respectively.

B. *MIRU-VNTR typing*

MIRU is a homologous 46-100 bp DNA sequence tandemly repeated and dispersed within intergenic regions of MTBC genome [16], [17]. Among 12-loci, 15-loci and 24-loci MIRU pattern analysis formats, we used 12-loci MIRU patterns in this study for genotyping MTBC [18]. The 12-loci MIRU pattern consists of loci 154 / MIRU02, 580 / MIRU04, 960 / MIRU10, 1644 / MIRU16, 2059 / MIRU20, 2531 / MIRU23, 2687 / MIRU24, 2996 / MIRU26, 3007 / MIRU27, 3192 / MIRU31, 4348 / MIRU39, and 802 / MIRU40. The MIRU pattern of an MTBC strain is represented as a vector of length 12, where each entry indicates the number of repeats in the specified MIRU locus.

C. *Mutation of spoligotypes and repeats in MIRU loci*

Spacers in the DR region can be lost as a result of chromosomal rearrangement event, but not gained [10], [11]. As a result, mutation of spoligotypes is unidirectional and can only result in variations of type $1 \rightarrow 0$ at each locus. Therefore, similar to Camin-Sokal parsimony, mutation of spoligotypes is irreversible [19]. Moreover, in a one-step mutation event, one or more contiguous spacers can be deleted. We refer to the rule of irreversible mutation of contiguous spacers as the *contiguous deletion assumption* [10], [11].

Tandem repeats in a MIRU locus can either be lost or gained as a result of duplication or multiplication in a mutation event [12]. Therefore, mutations at MIRU loci are bidirectional, and can result in increments or decrements in the number of repeats. The

variations in number of repeats in MIRU loci are simulated using the stepwise mutation model [20], [21].

III. Methods

A. The dataset

The dataset comprises of 9336 unique MTBC strains as determined by spoligotype and 12-loci MIRU patterns, collected by the United States Centers for Disease Control and Prevention (CDC) from MTBC isolates of patients in the United States from 2004 to 2008 [22]. There are 2841 unique spoligotypes and 4648 unique MIRU patterns. The strains are labeled by major lineages: East Asian (Beijing), East-African Indian (CAS), Euro-American, Indo-Oceanic, *M. africanum* and *M. bovis*.

B. Most parsimonious forest generation

We used both spoligotypes and MIRU patterns to simulate the evolution of DR loci reflected in spoligotype changes. We assumed that convergent evolution is rare, and loss of spacers is irreversible. We used three distance measures for strain comparison to generate the most parsimonious forest, which is the spoligoforest of MTBC strains.

1) Assumptions: Mutations in the DR region involve deletion of contiguous spacers, and acquisition of additional spacers is not observed, which is known as contiguous deletion assumption [8], [10], [11]. We also hypothesize in our model that convergent evolution does not occur. This is in accordance with the observations in the MTBC genome, with rare exceptions of homoplasy [10]. Using this hypothesis, we select the set of most likely parents for each spoligotype in the spoligoforest.

2) Distance measures for strain comparison: We used three distance measures based on two biomarkers of MTBC. Let $\vec{s}$ be 43-bit binary vector representing the spoligotype of an MTBC strain, and $\vec{m}$ be 12-bit vector representing 12-loci MIRU pattern of an MTBC strain. The biomarkers of MTBC are in the following format:

- $\vec{s}_i \in \{0, 1\}$, where $i \in \{1, .., 43\}$
- $\vec{m}_j \in \{0, .., 15\} \cup \{s, t, .., z\}$, where $j \in \{1, .., 12\}$ [1]

We defined three distance measures based on spoligotypes and MIRU patterns: Hamming distance between spoligotypes, Hamming distance between MIRU patterns, and L1 distance between MIRU patterns. Given two spoligotypes $\vec{s}_i$ and $\vec{s}_j$, the Hamming distance between them is defined as the number of spacers that differ:

$$H_S\left(\vec{s}_i, \vec{s}_j\right) = \sum_{r=1}^{43} \mid \vec{s}_{ir} - \vec{s}_{jr} \mid$$

where $\vec{s}_{ir}$ represents the presence of spacer r of spoligotype $\vec{s}_i$. Similarly, the Hamming distance between MIRU patterns is defined as the number of MIRU loci with different number of tandem repeats:

$$H_M\left(\vec{m}_i, \vec{m}_j\right) = \sum_{r=1}^{12} \mid sign\left(\vec{m}_{ir} - \vec{m}_{jr}\right) \mid$$

[1]The letters $\{s, t, .., z\}$ correspond to repeats with an additional mutation of 7 to 1 repeats respectively. Therefore, to separate these repeat values from the ones with numeric representation, the number of repeats $\{s, t, .., z\}$ are considered equivalent to 107 to 100 repeats respectively.

where $\vec{m}_{jr}$ represents the number of repeats at MIRU locus r of 12-loci MIRU pattern $\vec{m}_i$. To highlight the difference in the number of tandem repeats at each MIRU locus, we also defined the L1 distance between MIRU patterns:

$$L_M\left(\vec{m}_i, \vec{m}_j\right) = \sum_{r=1}^{12} \mid \vec{m}_{ir} - \vec{m}_{jr} \mid$$

In the spoligoforest, each spoligotype is associated with one or more MIRU patterns. Therefore, we calculate the Hamming distance and L1 distance between the MIRU patterns of spoligotypes as the minimum of distance values between sets of MIRU patterns associated with the two spoligotypes.

3) Validation of the model with segregation accuracy: Based on the assumption of negligibly infrequent convergent evolution, the task of generating a mutation model of spoligotypes reduces down to finding a unique parent spoligotype for each spoligotype, if a parent exists. First, we use the contiguous deletion assumption to find a set of candidate parent spoligotypes which may be immediate ancestors of the child spoligotype. Second, we use the three distance measures defined above to find the most parsimonious forest. There are six possible permutations of these distance measures. We used segregation accuracy to find the one which leads to most parsimonious spoligoforest. Segregation accuracy is defined as the percentage of within-lineage mutation events:

$$S = \frac{\sum_{l_i = l_j} d_{ij}}{\sum d_{ij}}$$

where d_{ij} is an indicator of a deletion event in which parent spoligotype $\vec{s}_i$ mutates into child spoligotype $\vec{s}_j$, and l_i represents the major lineage of MTBC strains with spoligotype $\vec{s}_i$. Maximum segregation accuracy is attained when the distance measures are used in the following order to pick the only parent among possible candidate parent spoligotypes: Hamming distance between MIRU patterns (H_M), Hamming distance between spoligotypes (H_S), L1 distance between MIRU patterns (L_M). Finally, if there still exists multiple parents, a single parent is chosen at random from the set of parent candidates. The flowchart of steps to pick a single parent for each spoligotype is shown in Figure 2.

4) The algorithm: Based on the flowchart in Figure 2, we generate the most parsimonious spoligoforest using Algorithm 1. Among all candidate parent spoligotypes, we first pick the parent spoligotypes that conform to the contiguous deletion assumption. Then, we reduce the size of the candidate parent set based on maximum parsimony using three distance measures in the following order: Hamming distance between MIRU patterns, Hamming distance between spoligotypes, L1 distance between MIRU patterns. Finally, if there are multiple parents still, we pick the parent spoligotype at random. We used variations of this algorithm. If only spoligotyping is used, steps 4 and 6 are skipped in Algorithm 1. If only MIRU typing is used, step 5 is skipped in the algorithm.

C. Statistical analysis of power law distributions

We observed power law distributions in the topology of spoligoforests, and tested the goodness-of-fit of these distributions. Power law distributions are often observed in the topological and graph-theoretical attributes of biological networks [23]. However, there

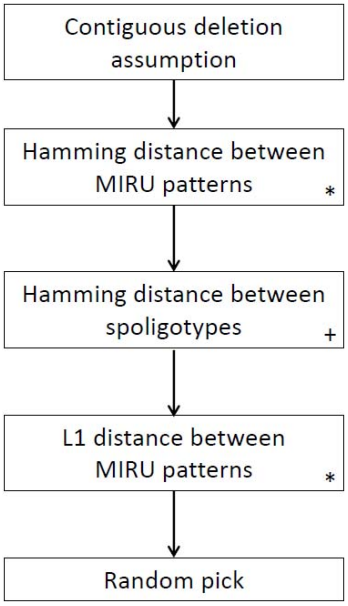

Figure 2: Flowchart of steps to pick the single parent for each spoligotype in the `MakeSpoligoforest()` algorithm using both spoligotypes and MIRU patterns. First, candidate parent spoligotypes are found based on the contiguous deletion assumption to ensure spacers in the DR region are only lost, but not gained. Then, to disambiguate the candidate parent spoligotypes, the Hamming distance between MIRU patterns, the Hamming distance between spoligotypes, and the L1 distance between MIRU patterns are used, resulting in minimum evolutionary change. Finally, if there are multiple parents still, a single parent spoligotype is picked from among the candidates at random. Spoligotype only variation of the algorithm skips the steps denoted with *, and MIRU pattern only variation skips the steps denoted with +.

Algorithm 1 `MakeSpoligoforest(StrainDataset)`

Input: `StrainDataset` with spoligotypes and MIRU patterns.
Output: Spoligoforest $G = (V, E)$, where node set V represents spoligotypes, and edge set E represents spoligotype mutations.

$E(G) = \emptyset$
for each node $s \in V(G)$ **do**
  Find the set of candidate parents P for node s using the contiguous deletion assumption.
  Find $P' \subseteq P$ with the minimum Hamming distance between MIRU patterns. Set $P = P'$.
  Find $P' \subseteq P$ with the minimum Hamming distance between spoligotypes. Set $P = P'$.
  Find $P' \subseteq P$ with the minimum L1 distance between MIRU patterns. Set $P = P'$.
  if $|P| > 1$ **then**
    Pick a node $p \in P$ at random
  end if
  Assign node p as the unique parent of node s.
  Add the edge e_{ps} from node p to node s.
    $E = E \cup \{e_{ps}\}$.
end for

is no single method widely accepted by scientific community for fitting power law distributions [24]. We adopt the method of analyzing power law distributions proposed by Clauset et al. [25]. According to this method, a power law distribution function is of the form:

$$p(x) = c\,x^{-\alpha}; \quad x \geq x_{min}$$

where α is the power law exponent, c is the normalization constant, and x_{min} is the lower bound for which the power law distribution holds. We also modified this function to fit a discrete power law distribution within a finite range. The method of maximum likelihood is used to estimate the exponent α. To find the lower bound x_{min}, the Kolmogorov-Smirnov statistic is used to measure the maximum distance between the cumulative distributions of the data and fitted power law model. The x_{min} value which minimizes this distance is selected as the lower bound. Finally, to test the goodness-of-fit of the power law distribution, we generate synthetic datasets from a true power law distribution using the same parameters, and compute the Kolmogorov-Smirnov statistic for each synthetic dataset relative to best-fit power law for that dataset. We calculate the *p*-value as the fraction of synthetic datasets for which Kolmogorov-Smirnov statistic is larger than the one observed for empirical data. If $p \geq 0.1$, then the power law distribution is a plausible fit to the data.

IV. Results

We generated the most parsimonious spoligoforest of the CDC dataset using Algorithm 1. The spoligoforest of the CDC dataset is shown in Figure 1. The spoligoforest shows the spoligotypes of MTBC strains, and a putative history of mutation events in the DR region reflected in spoligotype changes. Each node in the spoligoforest represents a set of MTBC strains with a distinct spoligotype. Each edge represents a potential mutation from parent spoligotype to child spoligotype. There are 2841 nodes and 2562 edges in the spoligoforest. 2547 of these edges represent a mutation event within the same major lineage, and the segregation accuracy is 0.9941. Among 2841 nodes, 232 of them are orphan nodes, represented by nodes with no parent or child node.

We compare the mutation model to existing mutation models of spoligotypes and verify that our mutation model leads to as many within-lineage mutation events as that of other mutation models. We observed interesting patterns on the topological properties of CDC spoligoforest and variations in DR loci. First, the number of descendant spoligotypes follows a power law distribution. Second, the mutation length frequency does not follow a power law. This contradicts the result of Reyes et al. that the mutation length frequency of spoligotypes follows a Zipf model [8]. Third, the number of deletion events at each contiguous DR loci follows a bimodal distribution. Based on this multimodal distribution, spacers 13 and 40 are identified as hotspots, and spacer 34 is the change point which is rarely exposed to mutation due to lack of spacer rather than low mutation rate.

A. Comparison to existing mutation models

Various models are used to generate a putative mutation history of spoligotypes. The Zipf model by Reyes et al. is based on their observation that length of unambiguous mutation events follows a Zipf distribution, and they generate spoligoforests based

Model	Segregation accuracy	# Isolated nodes	# Mutation events
Zipf model [8]	0.9921	235	2562
`MakeSpoligoforest()` (Spoligotyping)	0.9906	230	2562
`MakeSpoligoforest()` (MIRU typing)	**0.9941**	233	2562
`MakeSpoligoforest()` (Spoligotyping and MIRU typing)	**0.9941**	232	2562

Table I: Comparative analysis of mutation models of spoligotypes. All four models lead to high segregation accuracy, while `MakeSpoligoforest()` using both spoligotyping and MIRU-VNTR typing results in slightly higher segregation accuracy and maximizes the number of within-lineage mutation events, but the differences are not statistically significant.

on this probabilistic model [8], [26]. We used an independent biomarker, MIRU, to disambiguate the possible ancestors for each spoligotype. We compared the Zipf model to our model, `MakeSpoligoforest()` algorithm, with three variations: using spoligotyping only, using MIRU-VNTR typing only, and using spoligotyping and MIRU-VNTR typing as shown in Figure 2 in combination with the original version of the algorithm described in the methods section. Table I shows the statistics of the resulting spoligoforests based on these four models. In the spoligoforests of all four models, there are 2562 mutation events, represented by edges in the spoligoforest. Isolated nodes in the spoligoforests are the nodes with no parent or child spoligotype, so their ancestor and descendant spoligotypes are unidentified from the mutation history. The number of isolated nodes slightly differs, but they are close. Segregation accuracy is the highest in the spoligoforest based on `MakeSpoligoforest()` algorithm using both spoligotyping and MIRU-VNTR typing, and equal to the segregation accuracy of the spoligoforest generated using the variation of `MakeSpoligoforest()` algorithm with MIRU typing. However, the difference between the segregation accuracy of different mutation models is not statistically significant. This validates that simpler spoligoforest models based only on the length of deletion can be used with little or no degredation in the results.

B. Number of descendant spoligotypes

Each spoligotype can have at most one parent spoligotype, and any number of child spoligotypes, assuming convergent evolution does not occur. A single mutation event in the DR region results in a new child spoligotype. The number of immediate descendant spoligotypes for each spoligotype depends on the number of spacers present in the DR region, which is equivalent to the copy number of the spoligotype. In theory, the copy number of a spoligotype can range from 0 to 43, but not all spoligotype representations were observed in the dataset we analyzed. Let d_i be the number of descendants of the spoligotype represented by node s_i. Figure 3 shows the cumulative distribution of descendant spoligotype count frequency $P(D \geq d)$ on a log-log plot. We used the power-law fitting procedure introduced by Clauset et al. [25] to test whether the data follows a power law distribution. Table II shows the power law distribution function fit to the number of descendant spoligotypes. The power law distribution holds for all spoligotypes with $d \geq 2$ children spoligotypes. Based on the Kolmogorov-Smirnov test, the p-value of 0.6330 is larger than 0.1, which suggests that power law is a plausible fit to the number of descendant spoligotypes. The power law observation is based on two facts: 1) the higher the copy number of the spoligotype, the more descendants it can have, 2) the number of descendant spoligotypes increases due to the assumption of no convergent evolution, which leads to more genetic diversity. The number of descendants of a spoligotype can also be interpreted as the number of one-step deletion events that lead to new spoligotypes.

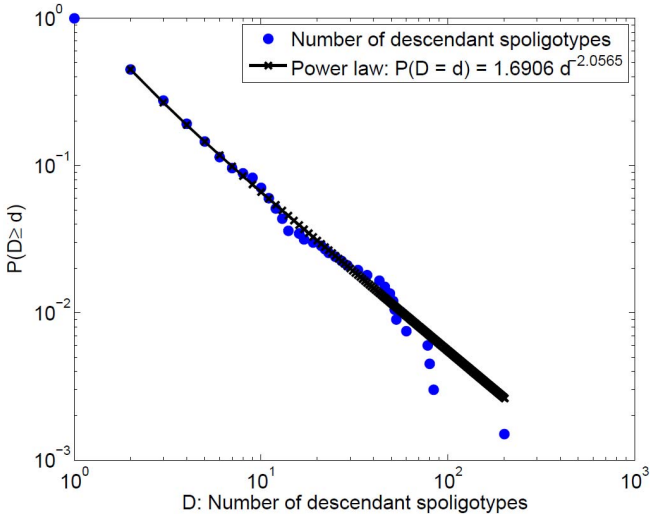

Figure 3: Number of descendant spoligotypes follows a power law distribution, which holds for spoligotypes with $d \geq 2$ children spoligotypes.

C. Mutation length

Mutation length is defined as the number of contiguous spacers deleted in a mutation event. Reyes et al. used only unambiguous deletion events in cluster-graphs, and observed that mutation length frequency follows Zipf distribution [8]. According to the contiguous deletion assumption, at each mutation event, a set of contiguous spacers can only be deleted, but not gained [10], [8]. In theory, mutation length can range from 1 to 43. Based on the putative mutation history of spoligoforest for the CDC dataset, we checked if a power law distribution is a plausible fit to mutation length frequency. We used the same procedure introduced by Clauset et al. to test whether the mutation length follows a power law distribution [25]. Let l_{ij} be the length of mutation event from node s_i to node s_j. Figure 4 shows the cumulative distributions $P(L \geq l)$ of two candidate power law distribution fits to the mutation length on a log-log plot. Table II shows two power law distribution function fits to mutation length, one in the range $[1, \infty]$, and one in the range $[1, 43]$. The first power law distribution holds only for the mutation events with length $l \geq 8$. Among all 2562 mutation events represented by edges in the spoligoforest, only 263 of them, which constitutes 10.27% of all mutation events, are of length $l \geq 8$. Therefore, power law distribution does not fit most of the

Attribute	Power law distribution function	Domain	p-value	Support for power law
D: Number of descendant spoligotypes	$P(D = d) = 1.6906\, d^{-2.0565}$	$d \geq 2$	0.6330	Good
L: Mutation length	$P_1(L = l) = 152.9498\, l^{-3.1020}$	$l \geq 8$	0.0020	None
	$P_2(L = l) = 0.6357\, l^{-2.0623}$	$1 \leq l \leq 43$	0	None

Table II: Candidate power law distributions and their goodness-of-fit test results based on Kolmogorov-Smirnov test. Number of descendant spoligotype frequency follows a power law distribution. On the other hand, the two power law distribution fits in the range $[8, \infty]$ and $[1, 43]$ are not plausible fits to mutation length frequency, as suggested by low p-values.

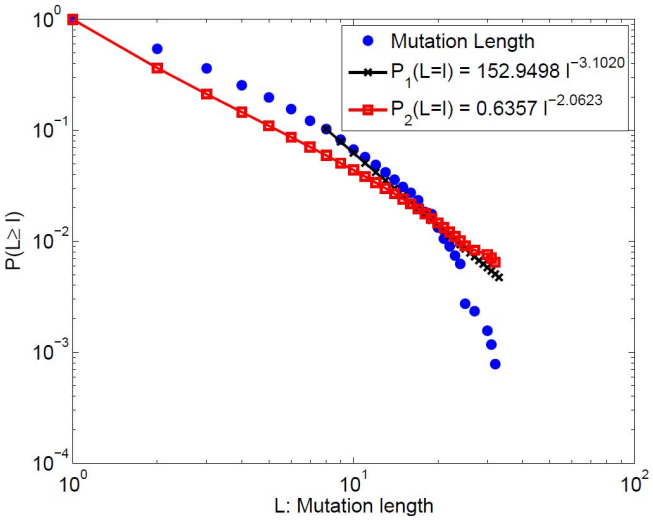

Figure 4: Mutation length frequency does not follow a power law distribution. The Kolmogorov-Smirnov test indicates that both power law distributions do not hold.

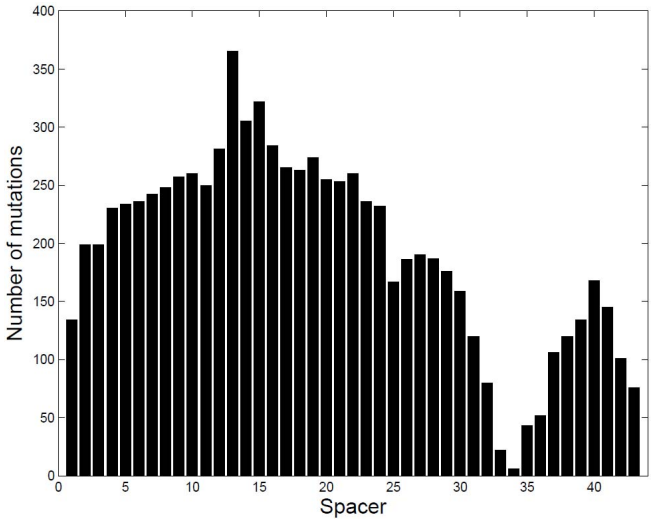

Figure 5: Number of deletions at contiguous DR loci follows a bimodal distribution. The two modes are spacer 13 and spacer 40, which are hotspots. The change point is spacer 34.

observed mutation events. Moreover, based on the Kolmogorov-Smirnov test, the p-value of 0.0020 is smaller than 0.1, which suggests that this power law distribution is not a plausible fit to mutation length. The second power law distribution fit in the range $[1, 43]$ has the probability mass function of the form used in Reyes et al. to fit the Zipf distribution [8]. The power law exponent is 2.0623, which is close to the power law exponent of 1.9962 found by the Zipf model in Reyes et al. using other datasets. However, the resulting p-value based on the Kolmogorov-Smirnov test is 0, and the second power law distribution is also not a plausible fit to mutation length. Therefore, mutation length does not follow a power law distribution, in contrast to the results in Zipf model built by Reyes et al. On the other hand, it is still accurate to claim that observed mutation patterns involve high numbers of short spacer deletions and small numbers of long spacer deletions. This is because mutation length depends on the number of contiguous spacers in the parent spoligotypes.

D. Number of mutations at each DR locus

Mutation events in the DR region lead to deletion of spacers. Spacers deleted in a mutation event can be counted to identify variations of the mutation rates in the DR region. Figure 5 shows the number of mutation events in which a spacer of each DR locus is deleted. Based on this figure, the number of deletions for each spacer follows a bimodal distribution, and the modes are spacer 13 and spacer 40. We call these DVR regions *hotspots*, or sites of increased observed variability. Spacer 34 is the change point in the bimodal distribution. This is due to lack of spacer 34 at DR region in most MTBC strains. In fact, out of 2841 spoligotypes in the CDC dataset, only 94 of them, which constitute 3.31% of all spoligotypes, have spacer 34 present in the DR region. Out of 9336 MTBC strains determined by spoligotype and 12-loci MIRU patterns in the dataset, only 192 of them, which constitute 2.06% of all MTBC strains, have spacer 34 present in the DR region. Therefore, the mutation rate is lowest at spacer 34 due to absence of spacer 34 in most MTBC strains in the CDC dataset. Two out of three principal genetic groups defined by Sreevatsan et al., PGG2 and PGG3, lack spacers 33 to 36, which is concordant with this observation [27]. In addition, 1971 spoligotypes out of 2841 in the CDC dataset are labeled with Euro-American lineage, which is characterized by the deletion of spacers 33-36 [28], [29]. This bimodal separation of DR loci leads to accumulation of shorter deletions among mutation events, rather than observing longer deletions, which explains why the power law distribution is actually not a plausible fit to mutation length.

V. Discussion and Conclusion

We developed a new mutation model of tuberculosis spoligotypes using the variations in the DR region and MIRU patterns to disambiguate the ancestors of a spoligotype. Based on the contiguous deletion assumption and no homoplasy, and using three distance measures, we generated the most parsimonious forest

of spoligotypes. The resulting spoligoforest depicts a putative history of mutation events in the DR region. Given the spoligotype mutations, we analyzed the biological network of spoligotypes in terms of both network topology and number of mutations at each DR locus.

We compared our mutation model based on spoligotypes and MIRU patterns with its counterparts using spoligotyping only, MIRU typing only and with Zipf model [8]. The mutation model which incorporates both biomarkers results in the most parsimonious spoligoforest and maximizes within-lineage mutation events. The comparison showed that segregation accuracy values are high in all four models with no statistically significant difference in the results. Therefore, spoligoforests created using only spoligotypes and the Zipf model are very similar to spoligoforests determined by the additional independent biomarker MIRU-VNTR. This validates the spoligoforest algorithms based only on spoligotypes, showing that spoligotype only algorithms can be used to generate the spoligoforest when MIRU patterns are not present.

The number of descendants of a spoligotype is equivalent to the outdegree of the corresponding node in the spoligoforest. We tested and verified the hypothesis that the number of descendant spoligotypes follows a power law distribution. This is due to the fact that the higher the copy number of spoligotype, that is, the more spacers present in the DR region, the more spoligotypes can descend from it. In addition, the assumption of no homoplasy favors genetic diversity rather than convergent evolution. Mutation length is the number of spacers deleted in a mutation event. We tested and verified that mutation length does not follow a power law distribution, as opposed to the Zipf model for mutation of spoligotypes proposed by Reyes et al. [8]. However, it is still accurate to state that mutations in the DR region rarely involve long deletions and frequently involve short deletions.

We calculated the number of mutation events which resulted in deletion of spacer at each DR locus. The number of mutations at consecutive DR loci showed an interesting pattern of bimodal distribution. The two modes are spacer 13 and spacer 40, which are hotspots of variations in the DR region. The change point in the bimodal distribution is spacer 34. This is due to absence of spacer 34 in a large number of MTBC strains, rather than low mutation rate at DVR34, because this spacer might be deleted irreversibly at the beginning of DR evolution and it can not mutate further after being deleted. Two out of three principal genetic groups defined by Sreevatsan et al. and MTBC strains of Euro-American lineage lack spacers 33-36, which supports the claim that low number of mutations in DVR34 is due to lack of spacer 34 in most MTBC strains in the CDC dataset [27], [28]. Since most of the deletion events occur either on spacers 1-34, or spacers 34-43, resulting in accumulation of shorter deletions, longer deletions are not observed. Therefore, this bimodal distribution explains why mutation length does not follow a power law distribution.

Future work will involve analysis of other topological attributes of the spoligoforest, extension of the mutation model to use other biomarkers, and interpretation of clades grouped closely in the spoligoforest. The mutation model can be extended to include more biomarkers, e.g. RFLP, with corresponding distance measures for the additional biomarkers to be used in the algorithm which generates the spoligoforest. Analysis of connected components in the spoligoforest can also give more insight into segregation of major lineages or sublineages. In addition, each tree or subtree in the spoligoforest can be a group of genetically related MTBC strains not classified as a separate clade earlier. This mutation model can also be extended to other organisms genotyped by CRISPR profiles.

ACKNOWLEDGMENTS

This work was made possible by Dr. Lauren Cowan and Dr. Jeff Driscoll of the Centers for Disease Control and Prevention. We thank Dr. Aaron Clauset and Dr. Cosma Shalizi for providing the implementations of the methods to fit power law distributions. This work was supported by NIH R01LM009731.

REFERENCES

[1] W. H. O. Report, *Global tuberculosis control : epidemiology, strategy, financing*. Geneva : World Health Organization, 2009.

[2] B. Mathema, N. E. Kurepina, P. J. Bifani, and B. N. Kreiswirth, "Molecular epidemiology of tuberculosis: Current insights," *Clin. Microbiol. Rev.*, vol. 19, no. 4, pp. 658–685, 2006.

[3] A. Shabbeer, C. Ozcaglar, B. Yener, K. P. Bennett, "Web tools for molecular epidemiology of tuberculosis," *Infection, Genetics and Evolution*, in press, 2011.

[4] C. Ozcaglar, A. Shabbeer, S. 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.

[5] R. Brosch, S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L. M. Parsons, A. S. Pym, S. Samper, D. van Soolingen, and S. T. Cole, "A new evolutionary scenario for the *Mycobacterium tuberculosis* complex," *PNAS*, vol. 99, no. 6, pp. 3684–3689, 2002.

[6] S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell, "Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence," *Nature*, vol. 393, no. 6685, pp. 537–544, June 1998.

[7] M. M. Tanaka and A. R. Francis, "Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information." *Infec Genet Evol*, vol. 5, pp. 35–43, 2005.

[8] J. Reyes, A. Francis, and M. Tanaka, "Models of deletion for visualizing bacterial variation: an application to tuberculosis spoligotypes," *BMC Bioinformatics*, vol. 9, no. 1, p. 496, 2008.

[9] A. Grant, C. Arnold, N. Thorne, S. Gharbia, and A. Underwood, "Mathematical modelling of *Mycobacterium tuberculosis* VNTR loci estimates a very slow mutation rate for the repeats," *Journal of Molecular Evolution*, vol. 66, pp. 565–574, 2008.

[10] R. M. Warren, E. M. Streicher, S. L. Sampson, G. D. van der Spuy, M. Richardson, D. Nguyen, M. A. Behr, T. C. Victor, and P. D. van Helden, "Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data," *J. Clin. Microbiol.*, vol. 40, no. 12, pp. 4457–4465, 2002.

[11] J. D. A. van Embden, T. van Gorkom, K. Kremer, R. Jansen, B. A. M. van der Zeijst, and L. M. Schouls, "Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria," *J. Bacteriol.*, vol. 182, no. 9, pp. 2393–2401, 2000.

[12] J. F. Reyes and M. M. Tanaka, "Mutation rates of spoligotypes and variable numbers of tandem repeat loci in *Mycobacterium tuberculosis*," *Infection, Genetics and Evolution*, vol. 10, no. 7, pp. 1046 – 1051, 2010.

[13] E. R. Gansner and S. C. North, "An open graph visualization system and its applications to software engineering," *SoftwarePractice & Experience*, vol. 30, no. 11, pp. 1203–1233, 2000.

[14] J. Zhang, E. Abadia, G. Refregier, S. Tafaj, M. L. Boschiroli, B. Guillard, A. Andremont, R. Ruimy, and C. Sola, "*Mycobacterium tuberculosis* complex CRISPR genotyping: improving efficiency, throughput and discriminative power of 'spoligotyping' with new spacers and a microbead-based hybridization assay," *Journal of Medical Microbiology*, vol. 59, no. 3, pp. 285–294, 2010.

[15] 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," *Journal of Clinical Microbiology*, vol. 35, no. 4, pp. 907–914, 1997.

[16] P. Supply, J. Magdalena, S. Himpens, and C. Locht, "Identification of novel intergenic repetitive units in a mycobacterial two-component system operon." *Molecular Microbiology*, vol. 26, no. 5, pp. 991–1003, 1997.

[17] P. Supply, E. Mazars, S. Lesjean, V. Vincent, B. Gicquel, and C. Locht, "Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome," *Molecular Microbiology*, vol. 36, no. 3, pp. 762–771, 2000.

[18] P. Supply, C. Allix, S. Lesjean, M. Cardoso-Oelemann, S. Rusch-Gerdes, E. Willery, E. Savine, P. de Haas, H. van Deutekom, S. Roring, P. Bifani, N. Kurepina, B. Kreiswirth, C. Sola, N. Rastogi, V. Vatin, M. C. Gutierrez, M. Fauville, S. Niemann, R. Skuce, K. Kremer, C. Locht, and D. van Soolingen, "Proposal for standardization of optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat typing of *Mycobacterium tuberculosis*," *Journal of Clinical Microbiology*, vol. 44, no. 12, pp. 4498–4510, December 2006.

[19] J. Camin and R. Sokal, "A method for deducting branching sequences in phylogeny," *Evolution*, vol. 19, pp. 311–326, 1965.

[20] M. Kimura and T. Ohta, "Stepwise mutation model and distribution of allelic frequencies in a finite population." *PNAS*, vol. 75, no. 6, pp. 2868–2872, Jun. 1978.

[21] T. Wirth, F. Hildebrand, C. Allix-Bguec, F. Wlbeling, T. Kubica, K. Kremer, D. van Soolingen, S. Rsch-Gerdes, C. Locht, S. Brisse, A. Meyer, P. Supply, and S. Niemann, "Origin, spread and demography of the *Mycobacterium tuberculosis* complex," *PLoS Pathogen*, vol. 4, no. 9, p. e1000160, 09 2008.

[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] G. Pavlopoulos, M. Secrier, C. Moschopoulos, T. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. Bagos, "Using graph theory to analyze biological networks," *BioData Mining*, vol. 4, no. 1, pp. 10+, 2011.

[24] G. Lima-Mendez and J. Helden, "The powerful law of the power law and other myths in network biology," *Molecular BioSystems*, vol. 5, no. 12, pp. 1482–1493, 2009.

[25] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661+, February 2009.

[26] C. Tang, J. Reyes, F. Luciani, A. R. Francis, and M. M. Tanaka, "spolTools: Online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex," *Bioinformatics*, vol. 24, no. 20, pp. 2414–2415, 2008.

[27] S. Sreevatsan, X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam, and J. M. Musser, "Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination," *PNAS*, vol. 94, no. 18, pp. 9869–9874, 1997.

[28] 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*," *PNAS*, vol. 103, no. 8, pp. 2869–2873, 2006.

[29] C. Borile, M. Labarre, S. Franz, C. Sola, and G. Refregier, "Using affinity propagation for identifying subspecies among clonal organisms: lessons from *M. tuberculosis*," *BMC Bioinformatics*, vol. 12, no. 1, p. 224, 2011.