Knowledge-based Bayesian network for the classification of *Mycobacterium tuberculosis* complex sublineages

Minoo Aminian
Departments of Mathematical Sciences and Computer Science
Rensselaer Polytechnic Institute
Troy, NY-12180
aminim@cs.rpi.edu

Amina Shabbeer
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY-12180
shabba@cs.rpi.edu,

Kane Hadley
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY-12180
hadlek2@cs.rpi.edu

Cagri Ozcaglar
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY-12180
ozcagc2@cs.rpi.edu

Scott Vandenberg
Department of Computer Science
Siena College
Loudonville, NY-12211
vandenberg@siena.edu

Kristin P. Bennett
Departments of Mathematical Sciences and Computer Science
Rensselaer Polytechnic Institute
Troy, NY-12180
bennek@rpi.edu

ABSTRACT
We develop a novel knowledge-based Bayesian network (KBBN) that models our knowledge of the *Mycobacterium tuberculosis* complex (MTBC) obtained from expert-defined rules and large DNA fingerprint databases to classify strains of MTBC into fifty-one genetic sublineages. The model uses two high-throughput biomarkers: spacer oligonucleotide types (spoligotypes) and mycobacterial interspersed repetitive units (MIRU) types to represent strains of MTBC, since these are routinely gathered from MTBC isolates of tuberculosis (TB) patients. KBBN provides an elegant and simple way to incorporate existing widely accepted visual rules for MTBC sublineages into a classifier designed to capture known properties of the MTBC biomarkers. Unlike prior knowledge-based SVM approaches which require rules expressed as polyhedral sets, KBBN directly incorporates the rules without any modification. Computational results show that KBBN achieves much higher accuracy than methods based purely on rules, and than Bayesian networks trained on biomarker data alone.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning; I.5.1 [**Pattern Recognition**]: Models - *statistical*; J.3 [**Computer Application**]: Life and Medical Sciences.

General Terms
Design, Experimentation, Verification.

ACM-BCB '11, August 1-3, Chicago, IL, USA

Keywords
Tuberculosis, Bayesian Networks, sublineages, spoligotype, multiple interspersed repetitive units

1. INTRODUCTION
One-third of the world is infected with tuberculosis (TB). Molecular epidemiology now plays a crucial role in the tracking and control of TB. DNA fingerprinting methods have made it possible to distinguish between cases of recent transmission of TB and reactivations of latent infections. This has enabled the tracking of transmission routes and the timely identification of outbreaks. Thus, knowledge about the genotype of prevailing strains has revolutionized traditional approaches in the epidemiology of TB. Moreover, the predominance of certain strains or groups of strains in certain host populations has clearly been observed [1, 2]. Studies of the genetic and biogeographic diversity of the *Mycobacterium tuberculosis* complex (MTBC) have revealed differences in the virulence, immunogenicity, and drug-resistance of strains [3-5]. This has consequences in the development of control measures for TB. Analysis of the population structure also provides insights into the evolutionary scenario of MTBC.

Phylogeographic lineages and sublineages have been defined based on genetic similarities between strains and observed associations between groups of similar MTBC genotypes with host populations [2, 6-8]. A variety of molecular techniques including the analysis of phylogenetically informative single nucleotide polymorphisms (SNPs) and long sequence polymorphisms (LSPs) are used to genotype MTBC strains. Classification based on SNPs and LSPs is considered to be the gold standard [2, 9-11]. However, studies of such variations in DNA sequences of MTBC strains are not performed frequently for public health purposes. Spacer oligonucleotide typing (spoligotyping) and mycobacterial interspersed repetitive units - variable number of tandem repeats (MIRU) typing are two polymerase chain reaction (PCR)-based DNA fingerprinting methods routinely used in the United States for genotyping all identified culture-positive TB cases. Spoligotyping is based on the polymorphisms found in the direct repeat (DR) region of the

mycobacterial chromosome, while MIRU typing is based on the number of tandem repeats present at 12 to 24 identified loci distributed across the MTBC genome [12]. Large databases of spoligotypes have been collected, the most significant being SpolDB4 comprising 39295 strains observed worldwide [6]. These strains have been assigned sublineage labels using a mixed expert-based and bioinformatical approach derived from visual rules.

The visual rules are based on the identification of characteristic deletions of one or more adjacent spacers. Certain inferred mutations (deletions of blocks of adjacent spacers) in progenitor strains are considered to be lineage-defining. These deletions are conserved in all descendent strains since studies have shown that the mechanism of evolution observed in the DR region involves loss of spacers, and spacers are rarely gained [13]. Additionally, the existence of these sublineages have been independently verified by clustering based on spoligotype and MIRU types of strains [14, 15]. Therefore, while it has been established that strains of TB belong to distinct sublineages, the definitions of these sublineages based on spoligotypes are not clear. The visual rules for a sublineage are generalizations of spoligotype patterns that belong to that sublineage. However, directly applying visual rules to spoligotype patterns can lead to multiple assignments of sublineage labels since spoligotype patterns may match patterns prescribed by more than one rule, and sometimes spoligotype patterns do not exactly match the patterns specified by any rule. This is an inherent limitation of a rule-based system – wherein rules need to be broad enough to capture general patterns, but narrow enough to delineate classes. Additionally, spoligotyping is based on polymorphisms in a single locus, the DR region, and therefore has the potential for convergent evolution. Relying on specific subsequences within the spoligotypes for the study of genetic diversity is hence error-prone.

This paper presents a hierarchical probabilistic graphical model, the knowledge-based Bayesian network (KBBN), that encodes the knowledge of MTBC obtained from expert-defined rules and large databases of DNA fingerprint data to classify strains of MTBC into sublineages. Expert knowledge is modeled in the top level of variables representing the rules. The middle level variables represent the class and the lower level represents various MTBC biomarkers – spoligotypes and MIRU types. These variables model the known properties of spoligotypes and MIRU repeats, as well as their mechanisms of mutation. The structure of the KBBN allows the knowledge base captured in the visual rules to be easily and simply incorporated into the learning method, while overcoming the limitations of using only specific deletions as specified by visual rules to decide the sublineage. Moreover, the incorporation of advice provides additional benefits in performance. The reasoning for any decision made by the KBBN is evident to users; the probability gives a quantitative estimate of the confidence.

Other approaches to incorporating advice in the form of rules has been shown to improve discriminative learning models of MTBC major lineages and other problems [16]. However, those methods are limited to rules expressed in less-intuitive polyhedral form. The proposed knowledge-based Bayesian network method allows the existing visual rules to be incorporated with no modification resulting in improved classification of sublineages over the predictions made with the visual rules or Bayesian Networks alone. Also unlike visual rules, the flexibility offered by the KBBN enables it to handle sublineages with no known rules.

2. BACKGROUND

2.1 DNA fingerprinting

Two frequently used DNA fingerprinting methods for the genotyping of MTBC strains are spoligotyping and MIRU typing. Because of their portable data format and reproducibility, these two fingerprinting methods have become the standard for individual strain identification for the purpose of TB control and tracking. Isolates from almost every TB patient in the United States are genotyped by these two methods. This has enabled the creation of large reference databases. We describe the two techniques here briefly, and mention key properties that were exploited in the modeling of variables for the design of the Bayesian network.

2.1.1 Spoligotyping

Spoligotyping is a PCR-based reverse hybridization technique that exploits polymorphisms in the DR region to distinguish between strains [17]. The DR region contains 36-bp repeats interspersed with up to 43 non-repetitive 31-41 bp length sequences called spacers. The spoligotype of a strain is represented as a 43-bit long binary string, with a 0 representing absence and 1 representing presence of a spacer sequence. A key fact about the evolution of spoligotypes is that once a spacer is lost, it is extremely unlikely to be regained. It is hypothesized that spoligotypes evolve by deletion of a single or multiple contiguous direct repeats (DRs), whereas insertion of DRs is very unlikely. [13, 18].

2.1.2 MIRU – Variable Number of Tandem Repeat (VNTR) Typing

MIRU-VNTR typing is a VNTR analysis bacterial typing scheme that provides a high-throughput reproducible method for molecular typing of MTBC. MIRU is a 46-100 bp DNA sequence dispersed within the intergenic regions of the MTBC genome as tandem repeats. MIRU typing is based on the number of repeats observed at certain identified polymorphic loci [19]. These loci are dispersed throughout the MTBC genome and are independent. The degree of discrimination between strains depends on the number of loci used. Twelve loci of MIRU are used in this study: MIRU locus 2677/MIRU24, and the following set of loci henceforth referenced as MIRU1: 154/MIRU02, 580/MIRU04, 960/MIRU10, 1644/MIRU16, 2059/MIRU20, 2531/MIRU23, 2996/MIRU26, 3007/MIRU27, 3192/MIRU31, 4348/MIRU39, and 802/MIRU40. MIRU typing has higher discriminatory power than spoligotypes, therefore especially when used in conjunction with spoligotypes, MIRU typing provides a powerful method for identification of strains [20].

2.2 Bayesian Networks

A Bayesian network (BN) is created to predict the 51 sublineages. A BN is a graphical representation of a probability distribution. Formally speaking, a BN is a directed acyclic graph $G(N,E)$ consisting of a set of nodes $X = \{x_i \mid x_i \in N\}$ to represent the variables and a set of directed links to connect pairs of nodes. Each node has a conditional probability distribution that quantifies the probabilistic relation between the node and its parents such that for a network of *k* nodes:

$$P(x_1, x_2, ..., x_k) = \prod_{i=1}^{k} P(x_i \mid parents\ (x_i))$$

Therefore, one can compute the full joint probability distribution from the information in the network. In other words, a well-represented Bayesian network can capture the complete nature of the relationship between a set of variables.

3. PRIOR BAYESIAN NETWORKS

SPOTCLUST was the first generative model used for analysis of MTBC sublineages [15]. SPOTCLUST uses mixture models based on spoligotypes to identify strain families of MTBC. The SPOTCLUST Bayesian Network models the asymmetric evolution of spacers using a Bayesian Network with "hidden parents" [15]. The hidden parents of a lineage generate the members of the lineage. They capture evolution of spoligotypes without generating the full phylogeny. A spacer in the hidden parent may be lost with small probability. A spacer that is absent in the parent is almost never gained. The design models the evolution mechanism of the DR region, allowing the Bayesian network to capture the deletions that are known to characterize spoligotype lineages. The hidden parent technique of SPOTCLUST is used for the spoligotype parts of the KBBN model.

The Conformal Bayesian Network (CBN), shown in Figure 1, is another generative model for analysis of both spoligotype and MIRU type data for MTBC strains [21, 22]. CBN captures domain knowledge about the properties of spoligotypes and MIRU and uses this information to classify MTBC strain genotyping data into major lineages. The value of locus MIRU24 generates the lineage, which in turn, determines the number of repeats in the remaining MIRU loci. Thus, patterns in the occurrences of repeats at each locus for each lineage are captured. The lineage also generates the hidden parents of the lineage which in turn generate the spoligotype spacers. CBN reflects the known mechanisms of evolution of the spoligotypes and MIRU. With rare exceptions, ancestral strains have 2 or more repeats at MIRU24. Thus the top-level variable, M_{24}, indicates whether MIRU24 is less than two (indicating modern lineages with high probability) or at least two (indicating ancestral lineages with high probability).

We tried using CBN to classify MTBC genotyping data into sublineages. But using the single rule: MIRU24 $\geq 2 \rightarrow$ancestral, as in the original CBN was not enough to generate a good model. KBBN grew out of the effort to incorporate all of the visual rules available from SpolDB4 [6], the Fourth International Spoligotyping Database.

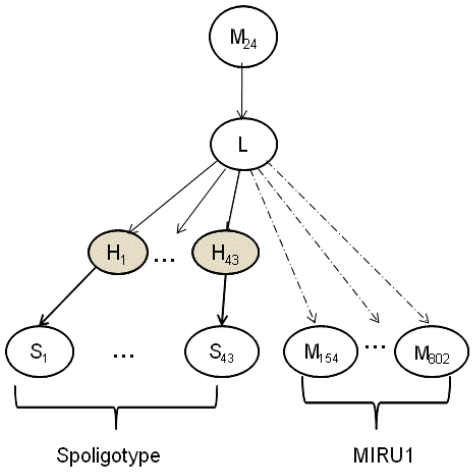

Figure 1. The Conformal Bayesian network uses a single rule based on the number of repeats at the MIRU24 locus as the first-level of a hierarchical Bayesian network. It uses the 43 spacers as features. The shaded nodes refer to hidden variables that model the fact that spacers are lost but rarely gained. In addition, the number of repeats at MIRU loci may be used. The nodes pointed to with dotted lines are not used for prediction. CBN uses a rule based on the value of locus MIRU24 to predict the major lineage with high accuracy.

4. VISUAL RULES

The SpolDB4 visual rules for MTBC sublineages are based on spoligotype patterns. Forty-eight visual rules are given in [6]. A sample of these rules is presented in Figure 2. Each line corresponds to a rule. The underlined portions of the spoligotype must match exactly while the portions not underlined can take any value. Note that in [6], the rules are expressed using the octal coding of spoligotypes; here we express them in binary for simplicity. While these rules establish characteristic patterns for sublineages of MTBC, they are not exclusive and in some cases overlap. In practice, a precedence or order is introduced over the

Sublineage	Binary Spoligotype Pattern	Rule
LAM2	■■□■■■■■■■□■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	**110**11111111**011**111**110000**11111111**0000**1111111
LAM5	■■■■■■■■■■■□■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	11111111111101111**110000**111111**110000**1111111
LAM9	■■■■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■■	1111111111111111111**110000**111111**110000**1111111
T1	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	1111111111111111111111111111111**110000**1111111
Spoligotype	11011111111101111111000011111111000011111111	Rules fired: LAM2, LAM5, LAM9, T1
	11111111111100111111000011111111000011111111	Rules fired: LAM5, LAM9, T1

Figure 2. Visual rules for four sublineages LAM2, LAM5, LAM9, and T1 from SpolDB4 knowledge base of rules. The rule column represents characteristic patterns specified by the visual rules as underlined subsequences in the spoligotype patterns. All of these rules fire for the spoligotype 11011111111101111111000011111111000011111111, while three of the rules fire for 11111111111100111111000011111111000011111111.

rules using expert knowledge so that unambiguous sublineage predictions are generated. However, this precedence has not been published for sublineages and is up to the individual user of the rules. The precedence reported here was crafted by the authors by trial and error to achieve good overall results. Another complication is that some SpolDB4 sublineages have no associated rules.

Visual rules with precedence have been established for six **major** MTBC lineages [23]. A prior online knowledge-based support vector machine (SVM) approach combined these visual rules and precedence into a set of rules expressed in polyhedral form [16]. The method produced a high accuracy SVM using much less data. However, this elegant work has several practical limitations that we sought to overcome in this study. First, expressing rules and precedence as polyhedral rules can be challenging for a large number of rules. Second, the method works best with linear SVMs and linear SVMs do not capture the underlying complexity of the biomarkers and their mechanism of evolution. This can be overcome by using nonlinear SVMs (3 degree polynomial kernels work very well) but then incorporating the polyhedral rules becomes even more challenging. Third, the complexity of training increases with the introduction of rules. Thus, the proposed design of the KBBN has the following salient features:

- Incorporates rules easily without modification, and without imposing precedence.
- Models known properties of biomarkers and their mutation mechanisms.
- Provides an efficient training method for classes with and without rules.
- Achieves high prediction accuracy.

5. KNOWLEDGE-BASED BAYESIAN NETWORK

The Knowledge-Based Bayesian Network (KBBN) represented in Figure 3 is a novel hierarchical Bayesian network probability model for sublineage classification of MTBC. KBBN captures domain knowledge about the properties of spoligotype and MIRU and incorporates additional information provided by SpolDB4 rules to predict the class with high accuracy. The corresponding probability density function for the model, shown in Figure 3, is:

$$P(C, M, S_\Omega, R_\Psi) = \sum_H \left(\prod_{j\in\Omega} P(S_j \mid H_j) P(H_j \mid C) \prod_{i\in\Gamma} P(M_i \mid C) P(C \mid R_\Psi) P(R_\Psi)\right)$$

where the random variable C represents the sublineage class, the random variable $S_\Omega = \{S_j \mid j \in \Omega\}$ with $\Omega = \{1,...,43\}$ represents the spoligotype spacers, the random variable $M_\Gamma = \{M_i \mid i \in \Gamma\}$ $\Gamma = MIRU1$ represents the MIRU loci as indexed by their locus number, and finally $R_\Psi = \{R_k \mid k \in \Psi\}$ $\Psi = \{1,...,46\}$ represents the set of binary rules indicating whether or not each specific rule is fired. All the variables are assumed to follow a binomial or multinomial distribution.

As in [21], each MIRU locus except MIRU04 is modeled as a multinomial distribution with possible values 0, 1…8, and ≥ 9. MIRU04 can take on some additional values. Since the proportions of different classes are not equal and some copy numbers do not occur, we used Dirichlet smoothing with non-uniform priors. The reader can consult the description of CBN in [21] for full details of these portions of the model.

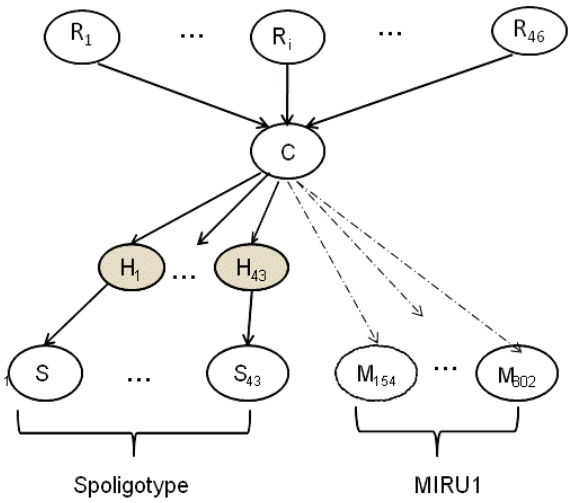

Figure 1. The KBBN uses multiple rules based on the presence of characteristic deletions at the first-level of a hierarchical Bayesian network. As with the CBN, it uses the 43 spoligotype spacers and/or number of repeats at MIRU loci as features. The shaded nodes refer to hidden variables that model the fact that spacers are lost but rarely gained. The dotted lines indicate that the MIRU1 variables can be dropped to create a spoligotype-only model.

For spoligotypes, we followed the SPOTCLUST model [15]. It captures the fact that spacers are lost but almost never gained, by introducing a variable for the unobserved hidden parent (H_j) and for each spacer S_j, both of which follow a binomial distribution. Given a 43-dimensional spoligotype S and its spacer position j, $S_j = 1$ if spacer is present, and $S_j = 0$ if spacer is absent. The probabilities of the spacer given the parent $P(S_j|H_j)$ are assumed to be known. As in [21], we considered the probability of losing a spacer as 10^{-1} and probability of gaining a spacer equal to 10^{-7}.

The KBBN assumes that the MIRU loci and the spoligotype hidden parents are conditionally independent given the sublineage. The MIRU loci are scattered throughout the chromosome of MTBC in locations away from the DR locus used for spoligotyping. Thus, the assumptions of independence between the MIRU loci, and between MIRU and spoligotype, are well supported biologically. The conditional independence assumption of spacers is a model simplification previously made in the SPOTCLUST BN model [15]. This conditional independence of the biomarkers in the BN model enables KBBN to conform to the set of available biomarkers without any expensive missing value computations. A spoligotype-only model can be created by simply dropping the MIRU1 variables. None of the genotyping variables in the BN are treated as unobserved except for the hidden parent spacers (which are always unobserved).

Using Bayes‟ rule, one can predict the sublineage for new data by determining the sublineage with maximum probability:

$$P(C \mid S_\Omega, R_\Psi) \propto \sum_H \left(\prod_{j\in\Omega} P(S_j \mid H_j) P(H_j \mid C) P(C \mid R_\Psi)\right)$$

6. EXPERIMENTAL RESULTS

We compared the performance of several variations of KBBN against five alternative methods. We constructed KBBN using spoligotype+MIRU data and spoligotype alone. Since determining the precedence of rules can be challenging, we implemented KBBN with overlapping rules (without precedence) and with precedence-imposed rules.

The five other methods are: precedence-imposed rules, rules without precedence, nonlinear SVM trained on spoligotype+MIRU, nonlinear SVM trained on spoligotype alone, and SPOTCLUST (a spoligotype-only BN with hidden parents). For SVM, the Weka package for multiclass SVM was used to construct the model [24]. The data was preprocessed by mapping the spoligotypes to -1,1 and normalizing the MIRU. The C parameter was selected by cross–validation. Ten-fold cross-validation of the training set parameters was used to select the degree of a polynomial kernel. Third degree polynomial kernels were found to work best over alternative kernels including Radial Basis Function (RBF) and linear kernels. In prior studies not reported here, we found that 3-degree polynomials perform best for SVM so we restrict presentation of results to that kernel. SPOTCLUST has proved to be one of the best of the non-knowledge BN that we tried so it was chosen as the base-line BN method.

The predictive accuracy of each model was measured by 10-fold stratified cross validation. The 10-fold training and testing sets were designed to be disjoint with respect to spoligotype and MIRU. This ensures there are at least ten unique genotypes (Spoligotype-MIRU pairs) in each sublineage considered. For methods using only spoligotype, the accuracy and F-measures are naturally slightly higher since the train and test sets may be overlapping.

6.1 Datasets

Two datasets were combined for use in this study. The first, *CDC*, was the data collected by the Centers for Disease Control, United States (CDC) as part of routine TB surveillance in the United States from 2004-09 consisting of 31,482 MTBC isolates genotyped by spoligotyping and 12-loci MIRU typing. The second dataset was the SpolDB4 dataset available in the online supplement of Brudely et al, 2006 [6] consisting of 1939 distinct spoligotypes labeled with 62 SpolDB4 sublineages. KBBN was trained on a dataset of 6778 records, each corresponding to a (spoligtype, MIRU, sublineage) triplet, obtained by joining the SpolDB4 and *CDC* datasets by spoligotype. In this study, every distinct MIRU and spoligotype pair is considered to be a unique genotype. We dropped the classes for which there were fewer than 10 records and also the sublineages that do not commonly infect human beings (e.g. PINI1 and PINI2). To keep small classes, we combined MANU1, MANU2and MANU3 into one class of MANU, and also combined AFRI1 and AFRI3 into one class of AFRI_1_3. Overall, we ended up with 51 classes. Data was preprocessed by adding an array of 46 binary values, each representing a SpolDB4 rule applied to each record. The value of the rule was set to 1 if the rule was fired and zero otherwise. If precedence is imposed, only the rule with highest precedence fires, otherwise multiple rules may fire.

Table 1. Comparison of F-values of KBBN, Rules, SVM, and SPOTCLUST based on out-of-sample 10-fold cross validation test results. Two sets of models were created for SVM and KBBN – one trained on spoligotype and MIRU and one trained on only spoligotype (*). The training and test folds are defined based on distinct spoligotype and MIRU types and are identical for all methods. The slightly higher F-values of the (*) models can be explained by the fact that some spoligotypes were repeated in both train and test sets.

	Model	Ave. F-value ± stdev
Sp+MIRU	No precedence KBBN	0.975 ± 0.05
	Precedence-based KBBN	0.967 ± 0.04
	SVM	0.952 ± 0.08
Rules	Rules without precedence	0.605 ± 0.26
	Rules with precedence	0.946 ± 0.20
Sp Only	No precedence KBBN*	0.981 ± 0.06
	Precedence-based KBBN*	0.979 ± 0.07
	SVM*	0.991 ± 0.01
	SPOTCLUST*	0.823 ± 0.20

6.2 Prediction Results

The average F-value estimated by 10-fold cross validation for the four KBBN variations and five alternative methods is provided in Table 1. KBBN is better than or not significantly different from the alternatives. Neither the purely rule-based methods nor BN without rules are satisfactory by themselves. Rules without precedence perform poorly because no prediction is made when multiple rules fire. Imposing precedence on the rules moves accuracy over 94% but still significantly below KBBN. The spoligotype-only BN method, SPOTCLUST, only achieves 82.3 % accuracy. KBNN is very competitive with the best SVM methods, but offers additional advantages.

We also studied the out-of-sample prediction probability of each model for each sublineage and provided the prediction probability distribution map of each model per sublineage as shown in Figure 4. Note, the two rule-based methods fail, because no rules exist for some sublineages, such as BEIJING-LIKE, and rules with no precedence overlap.

7. DISCUSSION

The KBBN benefits from both expert advice and large collections of DNA fingerprint data. It performs significantly better than the original SPOTCLUST Bayesian network trained using only spoligotypes. It also outperforms the „Rules-only" systems even after the incorporation of precedence.

Rule-based systems require the exact matching of spoligotypes with specified patterns. These specified patterns correspond to inferred mutation events (deletions of one or more adjacent spacers) that characterize sublineages. However, often spoligotypes match patterns prescribed by more than one rule, and are thus assigned multiple sublineage labels. The deletion of contiguous spacers is usually considered to be a single mutation event rather than the independent loss of spacers over time [13]. However, since spoligotyping is based on variations within a single locus, the DR region, there is a potential for convergent evolution of spoligotype patterns [18]. This is often the reason

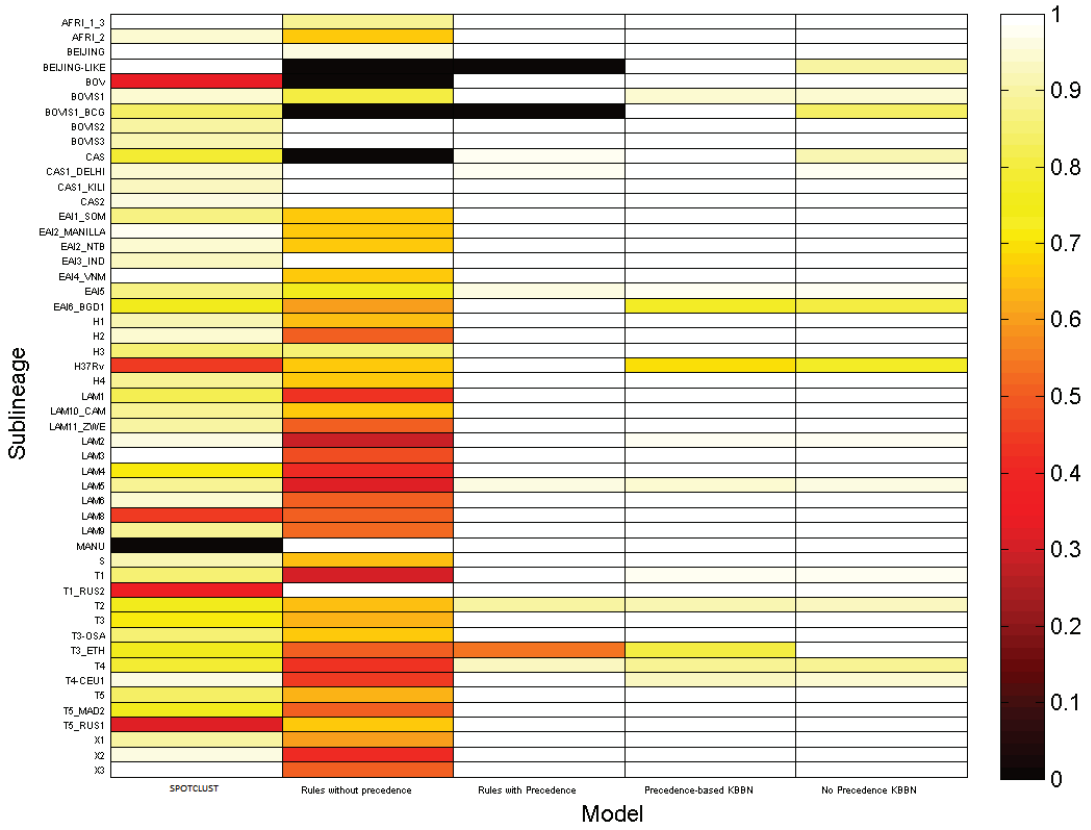

Figure 2. The heatmap represents the average F-value for the 51 lineages as determined by the 5 models: 1) SPOTCLUST (BN using spoligotypes alone) 2) Rule-based system without precedence 3) Rule-based system with precedence 4) Precedence-based KBBN trained on spoligotypes and MIRU 5) KBBN with no precedence trained on spoligotypes and MIRU. The KBBN models have the best performance as observed from the dominance of white squares indicating high precision and recall as captured by average F-measure.

cited as a limitation of spoligotyping for population genetic analyses [25]. It is necessary to use all available evidence from the data that may indicate a sublineage, not just the presence of a few contiguous deletions such as in the visual rules. The KBBN model is well-suited to this task because it incorporates the probability distributions of all spacers in addition to the presence of characteristic deletions specified by rules. The KBBN model is extensible to other biomarkers, such as SNPs and additional MIRU loci, as well as rules for these biomarkers, as observed in [21].

In contrast, rule-based systems incorporate precedence to mitigate the effect of multiple labels by checking for more stringent patterns first. This does improve the accuracy over a basic rule-based system without precedence, but completely excludes from consideration any other potential sublineage labels. Some spoligotype patterns do not exactly match the patterns specified by any rule. While probabilistic systems can handle such cases elegantly, these spoligotypes are typically assigned a "catch-all" label by a rule-based system with precedence. Thus, a deterministic rule based system is prone to some misclassifications.

Other successful strategies to mimic the behavior of rule-based systems, while mitigating its disadvantages by allowing for variations in definitive patterns, include nearest neighbor approaches as in [26]. However, these require the definition of appropriate distance measures and it is difficult to capture the dependence in the deletion of adjacent spacers in a distance measure. Nearest neighbor methods also require a comparison with every instance in the database in real time, in contrast to Bayesian networks.

These results indicate that high classification accuracy can be achieved using less data by the incorporation of domain knowledge in the form of rules. In this regard, KBBN is consistent with approaches from other methods. In [16], visual rules with precedence converted into polyhedral rules by experts led to better classification of MTBC major lineages by online SVM. KBBN improves on these approaches by allowing rules to be applied with or without precedence with no modification. Previously, in [21], it was shown that including a single rule based on the number of repeats at the MIRU24 locus that distinguish between „modern" and „ancestral" strains leads to improved accuracy.

8. CONCLUSION AND FUTURE WORK

KBBN is a high accuracy classifier for 51 MTBC sublineages that outperforms methods based on rules or Bayesian networks trained on data alone, and meets or beats the performance of nonlinear SVM models. As a general approach, KBBN has many attractive properties. It allows any type of rules to be incorporated into a Bayesian Network with little increase in the model and training complexity. Prior knowledge-based SVM required manipulation of the rules, models, data, and/or kernel [16, 27, 28]. In contrast to SVM, KBBN can produce explanations and probabilities of classes based on which rules were used and how they were affected by the rest of the KBBN.

KBBN can be readily extended to other learning tasks. It can perform unsupervised and semi-supervised learning by treating the class as unobserved for some training instances. This is an important property for TB sublineages since large unlabeled datasets exist and new lineages are being discovered. Incorporating patient characteristics into an unsupervised KBBN model can help it discover interesting host-pathogen groups.

In future work, we plan to evaluate KBBN as a general strategy for incorporating rules into Bayesian Networks on other domains and compare it with other strategies and knowledge-based learning methods. Also, we plan to create a more definitive KBBN model for MTBC sublineages based on the latest SITVIT sublineages, rules, and databases in conjunction with Institut Pasteur. The goal is to produce a publicly available web-based tool for sublineage classification to support TB control and research efforts.

9. ACKNOWLEDGMENTS

This work was made possible by and with the assistance of Dr. Jeffrey R. Driscoll and Dr. Lauren Cowan of the CDC. We would like to thank Dr. Nalin Rastogi and Institute Pasteur for their assistance. This work is supported by NIH R01LM009731.

10. REFERENCES

[1] Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W. and Small, P. M. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *P Natl Acad Sci USA*, 101, 14 (Apr 6 2004), 4871-4876.

[2] Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B. C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M. C., Hilty, M., Hopewell, P. C. and Small, P. M. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A*, 103, 8 (Feb 21 2006), 2869-2873.

[3] Gagneux, S. and Small, P. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis*, 7, 5 (2007), 328 - 337.

[4] Kato-Maeda, M., Bifani, P. J., Kreiswirth, B. N. and Small, P. M. The nature and consequence of genetic variability within *Mycobacterium tuberculosis*. *The Journal of Clinical Investigation*, 107, 5 (2001), 533-537.

[5] Malik, A. N. J. and Godfrey-Faussett, P. Effects of genetic variability of *Mycobacterium tuberculosis* strains on the presentation of disease. *The Lancet Infectious Diseases*, 5, 3 (2005), 174-183.

[6] Brudey, K., Driscoll, J., Rigouts, L., Prodinger, W., Gori, A., Al-Hajoj, S., Allix, C., Aristimuno, L., Arora, J. and Baumanis, V. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpoIDB4) for classification, population genetics and epidemiology. *Bmc Microbiol*, 6 (2006).

[7] Filliol, I., Driscoll, J. R., van Soolingen, D., Kreiswirth, B. N., Kremer, K., Valetudie, G., Anh, D. D., Barlow, R., Banerjee, D., Bifani, P. J., Brudey, K., Cataldi, A., Cooksey, R. C., Cousins, D. V., Dale, J. W., Dellagostin, O. A., Drobniewski, F., Engelmann, G., Ferdinand, S., Binzi, D. G., Gordon, M., Gutierrez, M. C., Haas, W. H., Heersma, H., Kallenius, G., Kassa-Kelembho, E., Koivula, T., Ly, H. M., Makristathis, A., Mammina, C., Martin, G., Mostrom, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Niobe-Eyangoh, S. N., Pape, J. W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M. L., Stauffer, F., Suffys, P. N., Takiff, H., Texier-Maugein, J., Vincent, V., de Waard, J. H., Sola, C. and Rastogi, N. Global distribution of *Mycobacterium tuberculosis* spoligotypes. *Emerg Infect Dis*, 8, 11 (Nov 2002), 1347-1349.

[8] Filliol, I., Driscoll, J. R., van Soolingen, D., Kreiswirth, B. N., Kremer, K., Valetudie, G., Dang, D. A., Barlow, R., Banerjee, D., Bifani, P. J., Brudey, K., Cataldi, A., Cooksey, R. C., Cousins, D. V., Dale, J. W., Dellagostin, O. A., Drobniewski, F., Engelmann, G., Ferdinand, S., Gascoyne-Binzi, D., Gordon, M., Gutierrez, M. C., Haas, W. H., Heersma, H., Kassa-Kelembho, E., Ho, M. L., Makristathis, A., Mammina, C., Martin, G., Mostrom, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Niobe-Eyangoh, S. N., Pape, J. W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M. L., Stauffer, F., Suffys, P. N., Takiff, H., Texier-Maugein, J., Vincent, V., de Waard, J. H., Sola, C. and Rastogi, N. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. *J Clin Microbiol*, 41, 5 (May 2003), 1963-1970.

[9] Baker, L., Brown, T., Maiden, M. C. and Drobniewski, F. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerg Infect Dis*, 10, 9 (Sep 2004), 1568-1577.

[10] Filliol, I., Motiwala, A. S., Cavatore, M., Qi, W., Hazbon, M. H., Bobadilla del Valle, M., Fyfe, J., Garcia-Garcia, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M. I., Leon, C. I., Crabtree, J., Angiuoli, S., Eisenach, K. D., Durmaz, R., Joloba, M. L., Rendon, A., Sifuentes-Osornio, J., Ponce de Leon, A., Cave, M. D., Fleischmann, R., Whittam, T. S. and Alland, D. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol*, 188, 2 (Jan 2006), 759-772.

[11] Gutacker, M. M., Smoot, J. C., Migliaccio, C. A., Ricklefs, S. M., Hua, S., Cousins, D. V., Graviss, E. A., Shashkina, E., Kreiswirth, B. N. and Musser, J. M. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics*, 162, 4 (Dec 2002), 1533-1543.

[12] Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H. and Roring, S. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J Clin Microbiol*, 44, 12 (2006), 4498 - 4510.

[13] Warren, R. M., Streicher, E. M., Sampson, S. L., van der Spuy, G. D., Richardson, M., Nguyen, D., Behr, A. A., Victor, T.

C. and van Helden, P. D. Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: Implications for interpretation of spoligotyping data. *J Clin Microbiol*, 40, 12 (Dec 2002), 4457-4465.

[14] Ozcaglar, C., Shabbeer, A., Vandenberg, S., Yener, B. and Bennett, K. P. A clustering framework for *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors. In *Proceedings of the IEEE International Conference on Bioinformatics & Biomedicine* (Hong Kong, 2010)

[15] Vitol, I., Driscoll, J., Kreiswirth, B., Kurepina, N. and Bennett, K. Identifying *Mycobacterium tuberculosis* complex strain families using spoligotypes. *Infect Genet Evol*, 6, 6 (2006), 491 - 504.

[16] Kunapuli, G., Bennett, K., Shabbeer, A., Maclin, R. and Shavlik, J. Online knowledge-based support vector machines. *Machine Learning and Knowledge Discovery in Databases* (2010), 145-161.

[17] Kamerbeek, J., Schouls, L., Kolk, A., vanAgterveld, M., vanSoolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R. and Goyal, M. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol*, 35, 4 1997), 907 - 914.

[18] Driscoll, J. R. Spoligotyping for molecular epidemiology of the *Mycobacterium tuberculosis* complex. *Methods Mol. Biol*, 551 (2009), 117-128.

[19] Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B. and Locht, C. Variable human minisatellite like regions in the *Mycobacterium tuberculosis* genome. *Mol Microbiol*, 36, 3 (2000), 762-771.

[20] Supply, P., Lesjean, S., Savine, E., Kremer, K., van Soolingen, D. and Locht, C. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. *J Clin Microbiol*, 39, 10 (Oct 2001), 3563-3571.

[21] Aminian, M., Shabbeer, A. and Bennett, K. A conformal Bayesian network for classification of *Mycobacterium tuberculosis* complex lineages. *BMC Bioinformatics*, 11, Suppl 3 (2010), S4.

[22] Aminian, M., Shabbeer, A. and Bennett, K. Determination of Major Lineages of *Mycobacterium tuberculosis* using Mycobacterial Interspersed Repetitive Units. *IEEE International Conference on Bioinformatics & Biomedicine* (2009).

[23] Shabbeer, A. a. C., L. and Driscoll, J.R. and Ozcaglar, C. and Vandenberg, S. and Yener, B. and Bennett, K. P. TB-Lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. *Unpublished* Manuscript (2011).

[24] Holmes, G., Donkin, A. and Witten, I. H. *Weka: A machine learning workbench*. IEEE, City, (1994).

[25] Hershberg, R., Lipatov, M., Small, P. M., Sheffer, H., Niemann, S., Homolka, S., Roach, J. C., Kremer, K., Petrov, D. A., Feldman, M. W. and Gagneux, S. High Functional Diversity in *Mycobacterium tuberculosis* Driven by Genetic Drift and Human Demography. *Plos Biol*, 6, 12 (Dec 2008), 2658-2671.

[26] Allix-Beguec, C., Harmsen, D., Weniger, T., Supply, P. and Niemann, S. Evaluation and strategy for use of MIRU-VNTR*plus*, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol*, 46, 8 (Aug 2008), 2692-2699.

[27] Towell, G. G. and Shavlik, J. W. Knowledge-based artificial neural networks. *Artif. Intell.*, 70, 1-2 (1994), 119-165.

[28] Lauer, F. and Bloch, G. Incorporating prior knowledge in support vector machines for classification: A review. *Neurocomputing*, 71, 7-9 (2008), 1578-1594.