A data-mining approach to spacer oligonucleotide typing of Mycobacterium tuberculosis

M. Sebban\(^1\), I. Mokrousov\(^2\), N. Rastogi\(^2\) and C. Sola\(^2,\ast\)

\(^1\)French West Indies and Guiana University, TRIVIA, Department of Mathematics and Computer Science, Campus Fouillole, 97159 Pointe-à-Pitre Cedex, Guadeloupe and
\(^2\)Unité de la Tuberculose et des Mycobactéries, Institut Pasteur de Guadeloupe, BP 484, F-97165 Pointe-à-Pitre Cedex, Guadeloupe

Received on February 9, 2001; revised on June 25, 2001; accepted on October 9, 2001

ABSTRACT
Motivation: The Direct Repeat (DR) locus of Mycobacterium tuberculosis is a suitable model to study (i) molecular epidemiology and (ii) the evolutionary genetics of tuberculosis. This is achieved by a DNA analysis technique (genotyping), called spacer oligonucleotide typing (spoligotyping). In this paper, we investigated data analysis methods to discover intelligible knowledge rules from spoligotyping, that has not yet been applied on such representation. This processing was achieved by applying the C4.5 induction algorithm and knowledge rules were produced. Finally, a Prototype Selection (PS) procedure was applied to eliminate noisy data. This both simplified decision rules, as well as the number of spacers to be tested to solve classification tasks. In the second part of this paper, the contribution of 25 new additional spacers and the knowledge rules inferred were studied from a machine learning point of view. From a statistical point of view, the correlations between spacers were analyzed and suggested that both negative and positive ones may be related to potential structural constraints within the DR locus that may shape its evolution directly or indirectly.

Results: By generating knowledge rules induced from decision trees, it was shown that not only the expert knowledge may be modeled but also improved and simplified to solve automatic classification tasks on unknown patterns. A practical consequence of this study may be a simplification of the spoligotyping technique, resulting in a reduction of the experimental constraints and an increase in the number of samples processed.

Contact: csola@pasteur.gp

1 INTRODUCTION
As a result of increased human population migrations, infectious disease spreading may have global consequences. To avoid this, it is needed to detect emerging outbreaks before they reach epidemic stage (Grein et al., 2000). In this context, tuberculosis remains the leading cause of death by an infectious disease and its control relies both on improvement of drug availability and diagnosis abilities. Among new diagnostic tools, which have led to a better understanding of global tuberculosis epidemiology, DNA fingerprinting and the building of genotyping databases represent a new and powerful way to analyze tuberculosis transmission (van Embden and van Soolingen, 2000). Nevertheless, when handling a large amount of data, especially binary data, the astonishing ability of human brain to intuitively separate noisy from significant information is most efficiently challenged by computers, that may easily learn and reproduce some human tasks such as classification or similarity analysis. The Tuberculosis agent belongs to the Mycobacterium tuberculosis complex that may be split into various subclasses. With overlapping yet distinct epidemiologies, these include M. tuberculosis sensu stricto, Mycobacterium bovis, Mycobacterium africanum and two other less investigated subspecies M. microti and M. canetti, which will not be further discussed here. Among DNA fingerprinting studies, spacer oligonucleotide typing or spoligotyping has been applied to characterize these subtypes and has gained increased international acceptance because it may be both rapid and easily applied as a first line discriminatory test (Kremer et al., 1999). It is based on the properties of the Direct Repeat (DR) locus of the M. tuberculosis complex genome, one of the best known polymorphic loci of this pathogenic agent (Hermans et al., 1991; Groenen et al., 1993; van Embden et al., 2000). It is also a suitable genetic model to study recombination since it shows an extensive strain-to-strain polymorphism and may be used both for molecular epidemiological studies (Kamerbeek et al., 1997) and for molecular evolutionary studies (Fang et al., 1999; Sola et al., 2001a) on

\(\ast\)To whom correspondence should be addressed.
tuberculosis. This locus is composed of multiple identical or nearly identical (differing by one or a few nucleotides) 36-base pairs (bp) DR copies, which are interspersed by short and nonrepetitive inter-DR spacer sequences (between 35 and 41 bp long). The precise physiological role—if any—of this locus remains unknown. It has been suggested that a similar locus found in the Archaea *Haloferax mediterranei* may be involved in replica partitioning (Mojica et al., 1995). The association of one DR and one spacer is designated Direct Variable Repeat (DVR), and strains may consequently differ by one or more discrete DVRs (Groenen et al., 1993). Alternatively, strains may sometimes also differ by long deletions of DR repeats, and IS6110-mediated transposition or homologous recombination are likely to be involved in such structural changes (Fang et al., 1999; van Embden et al., 2000). Nevertheless, the mechanisms that generate this diversity of repetitive structures are yet poorly understood. However it may be speculated that a combination of replication slippage, homologous recombination as well as insertion-sequence-mediated events are driving forces that shape its evolution. Based on an initial representation of 43 specific inter-DR spacers, spoligotyping, a PCR-based reverse cross-blot hybridization procedure, was invented in 1997 (Kamerbeek et al., 1997); spacers 20, 21, 33–36 were extracted from *M. bovis* BCG sequence (Hermans et al., 1991), whereas others were extracted from the *M. tuberculosis* H37Rv reference strain (Groenen et al., 1993; Cole et al., 1998). Spoligotyping permits to subtype *M. tuberculosis* complex strains, either directly on sputum specimen, on cultures or even on ancient histological specimen and bone extracts. A first global phylogenographical classification of *M. tuberculosis* complex by spoligotyping was recently attempted by our team and two new phylogeographical classes were defined, the East-African Indian and the Latin American and Mediterranean clades (EA-I and LA-M respectively), that harbor specific spoligotyping signatures (Sola et al., 2001a,b). We built a first database, called DB1, composed of 7352 strains split into 342 shared types (spoligotypes common to two or more isolates) and representative of more than 60 countries. Our main aim in the first part of this paper consisted in generating knowledge rules which would be useful for automatically classifying new profiles. Due to the intrinsic properties of human brain functions, e.g. global memorization of various shapes and synthetical way to treat information, human experts may have difficulties to analytically express their knowledge under the form of formal decision rules. This is particularly true while handling binary data, i.e. spoligotyping patterns in our case. Indeed, data mining methods appears to be better suited than human observation to obtain relevant knowledge rules permitting to classify patterns, a task fit for the Knowledge Discovery in Database (KDD) field. Even if an approach consisting in deriving simple rules for classification is not new, it has not yet been applied for spoligotyping of *M. tuberculosis*. The results obtained show that the number of spacers can be dramatically reduced for determining the clade (class) of a new profile. However, the large size of DB1, resulting in the presence of noisy data (typing errors, mislabeled instances, misclassified examples, overlapping between 2 clades), required to ‘clean’ the database. Thanks to an efficient data reduction technique it was achieved by selecting only the relevant prototypes. These methods have been developed during the last decade to combat noisy data in the modern, larger databases. Nevertheless, since Prototype Selection (PS) is not the theoretical subject of this article, interested readers are referred to a number of papers dealing with this topic (Hart, 1968; Gates, 1972; Aha et al., 1991; Wilson and Martinez, 1998). In this article, we applied on DB1 a recently published PS method (Sebban and Nock, 2000) which consists in keeping profiles representative of the probability of density of each clade. This results in the deletion of border examples (overlapping), and useless examples at the center of clades. The results obtained show that the new knowledge base may be dramatically reduced after PS without reducing the decision rule performances. Lastly, we also analyzed a second data base, called DB2, where 25 new spacers were added to the 43 original ones. 10 new spacer sequences discovered in an *M. bovis* strain (isolate 401) were published previously (Beggs et al., 1996). Recently, 41 new spacer sequences among which 15 were retained in our study, were also described (van Embden et al., 2000). This constitutes a new representation space for spoligotyping which now includes 43 + 25, i.e. 68 spacers. The goal of this experiment was to study the contribution of the 25 new spacers on DB2, rather than comparing knowledge rules deduced from DB1 and DB2. Actually, because of the experimental difficulties, DB2 contains for the moment only 323 profiles, resulting in non-comparable knowledge bases. Preliminary results obtained show that decision rules after the insertion of the 25 new spacers on DB2 are fewer in number than before. Finally, a correlation search between the spacers showed the existence of both negative and positive correlations between specific spacers. These results suggest that potential topological constraints on the DNA of the DR locus may directly or indirectly shape its evolution.

## 2 SYSTEMS AND METHODS

### 2.1 Oligonucleotide design, 68-dimensional spoligotypes

The original set of 43 spacers initially published (Kamerbeek et al., 1997) was recently modified (van Embden et al., 2000). Parameters of the spoligotyping technique
were kept unchanged. The first and second additional sets of oligonucleotides (25mer), contained 10 (set A) and 15 (set B) oligonucleotides and were respectively selected according to sequences published (Beggs et al., 1996; van Embden et al., 2000). In each case, the best 25mer was selected within the DR spacer sequence with the software ‘Primers for the Mac’ (v1.0a Apple Pi, Ashland, MA).

Preliminary experiments for each oligonucleotide set (12.5, 25, 50 and 100 pmol/150 µl for set A, and 12.5, 40, 100 pmol/150 µl for set B) including adequate positive and negative controls strains were performed to optimize hybridization conditions. Final membrane preparation was performed using these optimized concentrations. Oligonucleotide description and concentrations are available upon request to the corresponding author. DB1 is part of an ongoing population-based study which contained at the time 7352 profiles, split into 342 shared-types, and which were representative of about 60 countries. A limited version of this database and the sources of the data were published recently (Sola et al., 2001a). Seven modified-spoligotyping experiments were performed and the origin of the DNAs was as follows: Japan (n = 6), Russia (n = 92), USA (n = 5), caribbean (n = 95) and Italy (n = 12). A total of 116 profiles were retained for DB2 construction and mixed to the 207 previously published profiles (170 M. tuberculosis sensu stricto + 37 M. bovis as described in van Embden et al., 2000). The number of profiles in (DB2) totaled 323, including orphan patterns, i.e. profiles found only once.

2.2 Notations and clades

Since we applied machine learning algorithms, standard notations of this field for describing databases from a modeling point of view will be used. Let (x₁, y₁) be an instance of the database, where x₁ is a p-dimensional vector and y₁ a belonging class. p corresponds to the number of features characterizing the n instances of a Learning Sample (LS). The following equivalence list was established for the specific problem to be treated here:

—p is the number of spacers (features or descriptors) in each spoligotype of the databases. p = 43 for DB1 and p = 68 for DB2.

—n is the number of spoligotypes (n = 7352 for DB1 and n = 323 for DB2), i.e. the size of the LS.

—xᵢ is the p-dimensional binary vector (xᵢ¹, xᵢ², ..., xᵢᵖ) corresponding to the p values taken by each spacer (0 or 1) for the ith instance.

—yᵢ is the class of the spoligotype xᵢ. The 7352 instances of DB1 were clustered in 9 classes, a priori defined by the human expert, using previously published results (Kremer et al., 1999; van Embden et al., 2000; Sola et al., 2001a,b). The 7352 strains are labeled according to criteria described below. Although the human expert may both recognize and efficiently label many specific fingerprints ‘by eye,’ the literal expression of formal and exclusive knowledge rules compatible with automatic information treatment, is currently a challenging task. Given the presence of noisy data, we will see in this manuscript, that the a priori visual rules will be most efficiently challenged by the machine learning and data-mining approaches.

• Afri, for *M. africanaum*, (nAfri = 180 strains): this subclass of *M. tuberculosis* complex, as recently described in Viana-Niero et al. (2001) includes any profile where spacers 8, 9 and 39 are absent.

• T, (nT = 1590): this clade is to be split in future into various yet undefined families. It includes any strains where at least one of the spacers 1–30 is present, spacers 33–36 are simultaneously absent, spacer 31 is present, spacer 9 or 10 is present, and at least one of the spacers 21–24 is present.

• Beijing, (nBeijing = 1268): this clade (van Soolingen et al., 1995) includes any profile where spacers 1–34 are absent.

• EA-I, for East-African–Indian, (nEA1 = 907): this clade (Källenius et al., 1999; Sola et al., 2001a) includes any profile where spacers 29–32 and 34 are simultaneously absent, and at least one of the spacers 1–30 is present.

• Haarlem, (nHaarlem = 1034): this clade (Kremer et al., 1999) includes any strain where spacers 31, 33–36 are simultaneously absent, and at least one of the spacers 1–30 is present.

• LAM-1, LAM for Latin America and Mediterranean, (nLAM1 = 819): this clade (Sola et al., 2001a) includes any strains where spacers 21–24, and 33–36 are simultaneously absent, and at least one of the spacers 1–30 is present.

• LAM-2, (nLAM2 = 294): this clade is an attempt to define a new family. It includes any strain where spacers 9–10, and 33–36 are simultaneously absent, and at least one of the spacers 1–30 is present. As shown in Figures 1 and 2 subfamilies of LAM1 and LAM2 will be defined by SIPINA and were not defined by the expert. This discrepancy suggests that the visual rules defined by the expert can not be easily modelized by SIPINA because of spoligotypes harboring large deletions which currently jeopardizes classification of these families. Moreover recent yet unpublished

1 Descriptions of DB1 and DB2 are available upon request to the authors.
observations suggests that there is indeed more than 2, yet undefined subclades among the LAM strains.

- \( X, (n_X = 1186) \): this clade is currently found to be highly prevalent in some English-speaking countries (Soini et al., 2000, our own unpublished observations). It includes any profile where spacer 18 is missing, spacers 33–36 are absent, and missing spacer 18 may sometimes be linked to the absence of spacers 39–42.

- \( M. bovis, (n_{M. bovis} = 74) \): this well-defined subclass includes any strain where spacers 39–43 are simultaneously absent, and spacers 33–36 are simultaneously present.

Some of the clade or subclass definitions cited above may be considered as having been validated independently by other investigators (e.g. for \( M. africanum, \) Haarlem, Beijing, EA-I, \( M. bovis \)) whereas the description of others is either new and speculative (e.g. LAM-2, clade X) or remains undefined (clade T). Although the search of knowledge rules and definition of clades are distinct processes using different softwares, both processes are interlinked and results in one domain may boost the knowledge in the other one. We will see that this is indeed the case and that data-mining constitutes a powerful approach when handling genotyping databases.

2.3 Knowledge rules induced by decision trees

In part, induction algorithms were applied on the original LS in order to generate decision trees. These trees are built by splitting the sample into sub-samples, according to an optimized criterion (often an information measure) which is in our case the best discriminant spacer \( x^j \); each subset is then split according to the same strategy resulting into two new subtrees, and so on, until the information gain is sufficient. Once the tree is built, each path from the root to a leaf describes a decision rule (or knowledge rule), useful for understanding the model but also for labeling new examples. This type of model is very useful in fields in which the understandability of the deduced model is important. For instance, while it is not very crucial to understand what is the reasoning of a model specialized for the postal code recognition (in this case an artificial neural network can be very efficient), the understandability of an artificial model for a medical diagnosis system is obviously more important. This is the case for our problem. Decision trees are well suited, not only for providing very efficient classification models but also for describing the process from the input vector \( x_i \) (the spacers) to the predicted output label \( y_i \) (the class). In these experiments, we used the SIPINA for Windows software, developed at the ERIC laboratory of the University of Lyon 2 (http://eric.univ-lyon2.fr). This software collects the state-of-the-art tree induction algorithms, such as C4.5 (Quinlan, 1993), CART (Breiman et al., 1984), and SIPINA (Zighed and Rakotomalala, 2000).

2.4 Accuracy estimation by cross-validation procedure

In order to assess the predictive ability of a model, i.e. the determination of the belonging class of a new spoligotype, a cross-validation procedure is frequently used in the computer science and statistics fields (Efron and Tib-
shirani, 1993). Based on the results obtained on the LS, an unbiased estimation of the theoretical generalization accuracy of the model to the whole population is provided by this method. Then, the generalization accuracy is estimated from the information contained in LS by this cross-validation procedure. The principle is as follows: the original sample is divided into \( f \) folders. A predictive model (a decision tree, in our case) is built from \( f - 1 \) subsets during a learning stage, and tested on the remaining one. The same procedure is repeated for the \( f \) permutations. The estimation of the generalization accuracy is the average of the \( f \) accuracies computed at each stage of the cross-validation. Such an estimation method will be used in the following experimental section.

2.5 Prototype selection

Prototype selection was historically used to improve the efficiency of the \( k \)-nearest-neighbor classifier (Hart, 1968), which classifies an unknown instance according to a local vote by its \( k \)-nearest neighbors and according to a given distance function (often the Euclidean). Although its use was widely spread and encouraged by early theoretical results linking its generalization error to Bayes risk, this classifier suffers from several practical limitations (Breiman et al., 1984). Firstly, it is computationally expensive because it stores all the instances in memory. Secondly, intolerant of noisy instances, and of irrelevant attributes, it is sensitive to the chosen distance function. Pioneer works in PS firstly searched only solutions to solve the two first problems listed above (Hart, 1968; Gates, 1972; Aha et al., 1991). The third problem (intolerance of irrelevant attributes) is a matter of feature selection. PS algorithms usually use the Euclidean distance. This function is appropriate only if the attributes are numeric. Wilson and Martinez (1998) proposed new distance functions to handle numeric and nominal attributes.

In this paper, we used a recently proposed method Prototype Selection using Relative Certainty Gain (PSRCG; Sebban and Nock, 2000). We investigated the PS as an information preserving problem. Rather than optimizing the accuracy of a classifier, we built a statistical information criterion for Relative Certainty Gain (RCG) based on a quadratic entropy computed from the \( k \)-nearest-neighbor topology. From neighbors linked by an edge with a given instance, PSRCG computes a quadratic entropy by taking into account the label of each neighbor. From this entropy (which conveys a local uncertainty), it deduces a global uncertainty of the LS. While an instance deletion is statistically significant, PSRCG eliminates uninformative examples. Using \( k \)-nearest neighbors for building the neighborhood graph, PSRCG has shown a high ability for reducing the database and improving classification performances (Sebban and Nock, 2000). In the following section, we will use PSRCG as a data preprocessing process on the 7352 profiles, before constructing a new decision tree. By eliminating irrelevant strains, we aim at improving the current knowledge base induced for automatically classifying spoligotypes.

3 EXPERIMENTAL RESULTS

3.1 Knowledge discovery from DB1

In this section, we automatically generated knowledge rules and built a predictive model from DB1. We applied the C4.5 induction algorithm to generate the decision tree described in Figure 1 from the 7352 strains. As shown in this figure, each path from the root to a leaf describes a knowledge rule. For the nine defined clades, the problem may be modeled by 14 knowledge rules only. Each of the 7352 strains, a priori labeled by the expert, belongs to only one leaf. Each leaf describes a given class, according to the majority clade represented in the subset of these profiles. Actually, a given strain \( x_i \) may belong to a leaf for which the majority clade is not \( y_i \), i.e. the clade of \( x_i \). It means that some leaves may be pure, i.e. they contain examples of a unique clade, whereas others describe a class by a relative vote. In the latter case, the minority examples represent the learning errors of the model. A learning error may either represent misclassification by the expert or be due to biological processes such as recombination or gene flows which may lead to fuzzy classification boundaries. Indeed, a decision rule may not only be assessed according to its accuracy on the LS, but also by the number of examples which follow this rule. The accuracy of the rule is described in Figure 1 by the Confidence Interval (CI) of the proportion of correctly classified strains (value within each circle). It is actually more suited to describe the accuracy by a CI, which takes into account the number of strains which follow the rule, rather than by the simple success rate. A CI, with a \( \alpha \) risk at 5%, is defined as follows:

\[
CI = p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}
\]

where \( p \) is the success rate and \( n \) the number of strains. It explains why there is no CI for 100% success rates because in these cases, \( (1 - p) = 0 \). The number of examples that follow the rule is represented by the strain proportion of the clade in which this conjunction of spacers is found (value on each circle). In order to compare the performances of this automatic model (decision tree) with the expert’s knowledge base previously mentioned in Section 2.2, we decided to characterize the two approaches by the following criteria:

- the number of rules: it describes the complexity of the model. We take into account here the conjunctions of spacers, disjunctions corresponding to different paths in the tree;
the mean size of the rules (number of spacers from the root to the leaf): note that between two decision trees with same performances, we usually choose the smallest one;

- the success rate on the LS: the accuracy is obviously the most popular criterion for assessing the quality of a model;

- the number of different spacers used in the decision rule: for our specific problem, we aim at using the smallest number of spacers for simplifying the experimental processing.

According to these criteria, the expert’s knowledge base and the decision rules induced by C4.5 on DB1 are summarized in Table 1. We may also note that:

- The expert’s ability to label an unknown strain is real in many cases. However, it may not be so straightforward to generate simple knowledge rules in all cases. Moreover, a knowledge base should respect a set of constraints, particularly the non-overlapping between rules. This is not the case for the rules listed above. This explains the weak accuracy on average (76.6%) of the human rule in comparison with the knowledge base automatically deduced from DB1 with SIPINA (97.6%), which increases to (99%) after PS. The observation of Figure 1 shows that the a priori expert clustering is validated by knowledge rules with strong CIs in most instances. In other cases, and for lesser defined clades (e.g. T, X, or LAM2) the existence of secondary or tertiary rules together with new subsets (e.g. X1, X2) suggests that today’s global classification of tubercle bacilli clades based on spoligotyping results remains sub-optimal and is likely to evolve in the near future.

- Except for two spacers (37 and 38), the whole information (41 spacers) contained in the spoligotypes to discriminate the nine clades was used by the expert. On the contrary, only 13 spacers are used by the automatically deduced model from DB1 on the 7352 strains, and this with a higher performance. The selected spacers are: 8–10, 17–19, 22, 31, 33, 34, 36, 42, 43. In other words, presence or absence of these spacers may be considered as highly informative in the 43 dimension space for classifying profiles. Neighborhood relationships may indicate that all recombination mechanisms do not act similarly on each spacer. As an example, spacers 9 and 31 are well known hot spots for IS6110 transposition, an evolutionary driving force that may change the DR locus structure (van Embden et al., 2000; Filliol et al., 2000; Legrand et al., 2001; Benjamin et al., 2001).

- Finally, while an average of 25.4 spacers per rule is used by the expert for finding the label, only 5.8 are used by the induced decision tree. This result may get practical application in a near future to simplify the spoligotyping technique when definition of major spoligotyping clades will be improved and minimal extra genotyping information will be required to automatically label an unknown strain.

The learning accuracy on DB1 suggests that it is possible to build a predictive model, which would automatically classify any new profile. To assess the accuracy generalization of this model, we achieved a 5-fold cross-validation procedure on DB1. A mean success rate of 98% with a standard deviation of 0.8% was obtained.

3.2 Prototype selection from the 7352 strains

In this section, we used the PSRCG algorithm (for more details see Sebban and Nock, 2000). This PS algorithm is based on the optimization of an information measure (a quadratic entropy). It consists in deleting the irrelevant strains and globally improves the information in the representation space. Practically, the irrelevant strains are those located at the clade boundaries, where the uncertainty is likely to be high. Moreover, strains at the center of clusters (subsets of strains of a same clade in the representation space) are also weakly relevant for classification purposes. PSRCG is particularly suited in such a context, because it is not dependent on a learning algorithm and does not generate an inductive bias. In other words, the resulting subsets of prototypes may then be used by an induction algorithm to build decision trees. Among the 7352 original strains, our algorithm kept 4014 profiles only. From this subset of strains, called DB1sub, the decision tree using SIPINA was built. The new model is presented in Figure 2. As it was done for the knowledge base deduced from DB1, we computed the same criteria as above on DB1sub. Results are included in Table 1. We note that the deletion of noisy data results in the decrease of the number of rules (11 versus 14), and in an increase of the accuracy (99.0 versus 97.6%). Moreover, the number of spacers is dramatically reduced. Actually, nine spacers are sufficient to totally discriminate the 4014 strains (8, 9, 18, 19, 22, 23, 31, 36, 43), among them eight were already present in the decision tree built from DB1. In the new tree shown in Figure 2, a closer insight at the subsets of the LAM clades (designated LAM11, LAM12 and LAM13 in that tree), show that this splitting represents left(LAM12) or right(LAM11) entire deletions of the DR locus, whereas LAM13 represent ‘true’ LAM as defined above. The full significance of this result requires further investigations.

3.3 Contribution of the 25 additional spacers

The database DB2 was constructed by assembling previously published (n = 207) and our own in house gener-
Table 1. Results according to four performance criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Rules deduced from</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expert</td>
</tr>
<tr>
<td>Number of rules</td>
<td>17.0</td>
</tr>
<tr>
<td>Mean size of the rules</td>
<td>25.4</td>
</tr>
<tr>
<td>Accuracy</td>
<td>76.6</td>
</tr>
<tr>
<td>Number of spacers</td>
<td>41.0</td>
</tr>
</tbody>
</table>

The contribution of the 25 additional spacers may also be assessed through a statistical study. We performed a principal component analysis from the two datasets, in order to compare the part of the total variance explained by the two main factorial axes. While this percentage of the total variance was around 53% by mapping individuals of DB2 from a 43-dimensional space to a planar representation, the part of the variance explained by the two axes was improved by the addition of 25 new spacers. Actually, about 60% of the total variance was explained by such a representation which confirms the contribution of these new 25 spacers. Lastly, by using this data analysis method, a correlation circle between spacers was drawn (see Figure 5). In this figure, the correlation level between two spacers is measured by the value of the angle between two points. High positive correlation between spacers is shown by an acute angle. Inversely, a negative correlation is shown as an obtuse angle and no correlation between spacers by a right angle. Figure 5 shows a very high negative correlation between spacers 41 and 47. This means that the presence of the spacer 41 is almost always accompanied by the absence of the spacer 47, and vice versa. This knowledge is refuted only for 12 individuals among 323, which are all characterized by a large deletion in the locus, likely to be linked to IS6110-mediated deletion events (van Embden et al., 2000). In this context, it is striking that the presence of an inverted IS6110 copy in the Beijing clade occurs precisely between these two spacers. This very same DVR 47 also harbors point mutations in the DR repeat in all the M. bovis strains investigated so far (van Embden et al., 2000). We also found a positive correlation between spacers 1, 21, 28, 29, 31 and 35, but these spacers are independent of spacers 41 and 47. The significance of this correlation remains to be investigated.

4 CONCLUSION

In this paper, we applied data mining methods to generate knowledge rules for solving classification tasks from existing or specifically created databases. To the best of our knowledge, this study is a first attempt to automatically discover simple knowledge rules from spoligotyping data. Amazingly, the generated rules are different and simpler than those previously defined by the expert. A possible explanation of this phenomenon may come from the fact that decision trees use pruning to avoid overfitting (resulting in deep trees), whereas the expert takes into account all data as signals, i.e. without differentiating noisy from unnoisy data. This is especially true for spoligotyping, where $2^{43}$ (and now $2^{68}$) combinations are theoretically

minated experimental profiles ($n = 116$). Since M. africanum strains were not yet tested for 25 additional spacers, the results described in this section may unfortunately not be applied to M. africanum. Besides, since DB2 is yet limited in size, our aim does not consist in comparing knowledge discovered in DB1 and DB2, but rather in DB2 before and after insertion of the 25 new spacers. Actually, the comparison between the two representation spacers (43 and 68) must be achieved according to the same LS. We obtained two new decision trees which are respectively shown in Figures 3 and 4. In order to facilitate the comparison, we now used as spacer labels their new position in the spoligotype represented by 68 features (for instance, the spacer 21 is the 21st value in the 68-dimensional representation, and not the 21st in the 43-dimensional space). The renum-bering of the spacers was previously done by van Embden et al. (2000), because of the conserved order of the spacers among different isolates of M. tuberculosis. Consequently, the numbers reflect the genetic organization in the genome. Among the 11 rules, 7 are very relevant, and describe classes Haarlem, M. bovis, EAI1, Beijing, LAM1, T1, LAM2 almost perfectly. The 4 others are either weakly accurate (T2, X) or have weak influence (T3, EAI2). Two spacers and consequently related DVRs (31 and 47), have been shown to harbor point mutations in the DR repeat itself (van Embden et al., 2000). This is a striking result that could be explained by the presence/absence of a single nucleotide polymorphism in the various clades described. Comparing the first and the second decision trees, the two following observations may be done:

1. The number of knowledge rules is reduced in the 68-dimensional space. 11 rules (with 10 spacers) in the 43-dimensional space for discriminating the 323 profiles into the 8 classes were inferred, whereas 9 spacers and 10 rules for reaching the same objective were required in the 68-dimensional space. Theoretically, a smaller model is desirable because it will not suffer from the overfitting of the learning data, and will be more reliable for classifying new unknown individuals.

2. The reduction of the tree size is explained by the presence of the new spacer 1, which allows to totally discriminate the right side of the tree. On the other hand, spacers 26 and 43 were required in the first decision tree. Thus, the presence of spacer 1 now allows to remove an irrelevant rule.
likely. A direct consequence of these simpler rules may be a change in the laboratory experimental setup by eliminating the use of uninformative spacers. Another consequence may be that uninformative spacers may be downweighted or excluded when dealing with phylogeny reconstruction using parsimony principles. The described process could also allow to automatically classify *M. tuberculosis stricto sensu* into the 8 or 9 validated clades. This would represent a first level of sub-classification for epidemiological purposes, an application that deserves further investigation. Future investigations will intend to study other *M. tuberculosis* representations, such as the Variable Number of Tandem DNA Repeats (VNTRs) or the Mycobacterial Interspersed Repetitive Units (MIRUs; Frothingham and Meeker-O’Connell, 1998; Supply et al., 2000). Data-mining methods will undoubtedly allow to distinguish useful information from noisy data obtained by these markers.

**ACKNOWLEDGEMENTS**

This work was supported by the Délégation Générale au Réseau International des Instituts Pasteur et Instituts Associés and the Fondation Raoul Follereau. We are also very grateful to Dr C.Mammina, Italy, Dr B.Vishnevsky and Dr T.Otten, Russia and Dr H.Kasai, Japan, for providing some DNAs of *M. tuberculosis* clinical isolates.

**REFERENCES**


Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D. and Gordon,S.V. et al. (1998) Deciphering the biology of My-
**Mycobacterium tuberculosis** from the complete genome sequence. *Nature*, 393, 537–544.


