Machine learning for the prediction of antibacterial susceptibility in *Mycobacterium tuberculosis**

Katherine E. Niehaus[1], Timothy M. Walker[2], Derrick W. Crook[2], Tim E. A. Peto[2], and David A. Clifton[1]

***Abstract*— The prevalence of antibiotic resistance in pathogens is far outpacing our ability to develop new antibiotics. This necessitates the development of diagnostic tests that can determine bacterial susceptibility. For *Mycobacterium tuberculosis* (MTB), this is particularly urgent given that current methods for testing susceptibility take up to two months. The decreasing cost and time required for whole genome sequencing (WGS) offers the possibility of using genome-wide mutational patterns in bacterial DNA to determine drug susceptibility. However, the computational framework for taking advantage of this data has not yet been developed.**

This paper describes a machine-learning approach for predicting bacterial susceptibility from genomic data. The presence or absence of over 500 single nucleotide polymorphisms (SNPs) found in a dataset of 652 bacterial isolates from the Oxford University Hospitals NHS Trust and elsewhere in the UK were used as features for a number of classification algorithms. Susceptibility and resistance were defined based upon phenotypic growth patterns, and the results from the proposed machine learning method were compared to predictions based upon the presence of a set of known resistance-conferring mutations. Misclassified isolates were also examined for commonalities, revealing eleven potentially new resistance-conferring mutations. The prediction of drug susceptibility using the proposed approach was very promising. Classification accuracy of 93% was obtained for predicting resistance to isoniazid, a key first-line antibiotic drug for MTB. The proposed method was capable of particularly high sensitivity, ranging between 95-100% across the four drugs examined. There is great potential to further develop this framework to find new resistance-conferring mutations.

I. INTRODUCTION

Tuberculosis infects over one-third of the human population and claims over one million lives each year [1]. One of the primary challenges of clinical microbiology concerns the determination of the appropriate drugs for treating a bacterial infection. This is an especially important problem in light of the growing degree of bacterial resistance to common antibiotics. Multi-drug resistant *Mycobacterium tuberculosis* (MDR-MTB), defined as being resistant to the commonly-used antibiotics isoniazid (INH) and rifampicin (RIF), now accounts for about 3.6% of all active TB cases worldwide; there was nearly a doubling of diagnosed cases from 2011 to 2012 [2]. Although MTB and MDR-MTB are curable, it is difficult to estimate resistance in a timely manner.

The gold-standard method for determining MTB drug susceptibility is phenotyping through the proportion method, using Lowenstein-Jensen (LJ) solid media [3]. This method compares bacterial growth with and without the presence of an antibacterial drug. Crucially, the LJ proportion method can require up to two months to obtain results due to the slow growth of the bacteria in culture.

While phenotypic methods have long been the gold standard for susceptibility testing, nucleic acid-based methods have been clinically adopted more recently because of their faster turn-around time. Nucleic acid-based tests provide improvement upon slow phenotypic methods, but they are only capable of identifying resistant cases when a subset of known AR mutations are present, and they are only available for certain drugs [4].

A much more flexible approach lies in the incorporation of whole genome sequencing (WGS) into the clinical diagnostic pathway [1]. WGS differs from the nucleic acid methods presented above in that it reveals all of the single nucleotide polymorphisms (SNPs) in a given sample using a single test. Currently available sequencing methods require only about two days for complete processing (after growing the sample in culture for 7-10 days), with this time requirement only continuing to decrease [4]. WGS has already been found beneficial for epidemiological outbreak detection [5], but its utility for susceptibility testing has not yet been determined.

Many resistance-conferring mutations have been well-established for MTB. WGS could therefore be immediately useful for identifying the presence of all known AR mutations in a single test. However, the mechanism of resistance for many drug-resistant MTB isolates remains unknown. Between 10%-20% of INH-resistant isolates, for instance, lack a mutation in a known resistance gene [6]. WGS offers the opportunity to look across the entire genome for new SNPs that may be resistance-conferring. Given the number of SNPs that will be involved, machine learning techniques offer an analysis approach that can manage such large numbers of features. Therefore, WGS has the potential for even further clinical utility when examined through a machine learning framework. Here, an initial study was conducted to classify MTB isolates as being susceptible or resistant to various first-line drugs. In addition, the characteristics of those isolates that were misclassified based upon known AR mutations were examined more closely to provide clues towards promising mutations for future analysis.

*This work was supported by the UK Clinical Research Collaboration (UKCRC), the National Institute for Health Research (NIHR) Oxford Biomedical Centre, the Research Councils UK (RCUK), the Royal Academy of Engineering, and the Medical Research Council.

[1]K.E. Niehaus and D.A. Clifton are with the Institute of Biomedical Engineering, Department of Engineering Science, Oxford, OX1 3PJ, UK katherine.niehaus at eng.ox.ac.uk

[2]T.M. Walker, D.W. Crook, and T.E.A. Peto are with the Nuffield Department of Medicine, Oxford, OX3 7BN, UK

978-1-4799-2131-7/14/$31.00 ©2014 IEEE

II. METHODS

A. Phenotypic assessment

652 frozen sputum samples from 542 patients diagnosed with active MTB were obtained from University Hospitals, Birmingham, UK, the Oxford Hospitals NHS Trust, UK, and small clusters of cases from the surrounding regions. Bacterial MTB colonies were grown up for 1-3 weeks from each of the 652 samples. Phenotypic drug susceptibility for each drug was determined by performing an initial screen for resistance in liquid culture, which was then confirmed using LJ methods. The four common first-line drugs were examined: INH and RIF, documented above, as well as pyrazinamide (PZA), and ethambutol (EMB).

B. DNA sequencing

DNA was extracted from the bacterial cultures using the QuickGene DNA Tissue Kit S (Fujifilm, Japan). Standard lab and Illumina protocols were used to randomly fragment the DNA. Sequencing was performed with the Illumina HiSeq 2000. The standard sequencing protocol involves the analysis of each nucleotide base many times; each of these analysis iterations is termed a "read."

The alignment of the sequence was performed with respect to the reference MTB strain H37Rv (NCBI reference sequence NC_000962.2) using the commonly-employed software *Stampy*. Nucleotide bases were determined using standard filters, which take into account the sequencing and alignment quality, as well as the number of reads for each base. Single-nucleotide point mutations, nucleotide insertions, and nucleotide deletions were identified from 39 known resistance genes (out of the 3959 genes in the MTB genome) in comparison to the reference sequence. Some nucleotide bases cannot be called with confidence using this pipeline due to the low sequencing or alignment quality; these bases were considered to be mutations if the base with the highest number of reads did not agree with the reference.

C. "Direct Association" method of prediction

A catalogue of mutations already known to confer AR was compiled for each of the examined drugs; these mutations will be referred to as "known" AR mutations. Any SNP found in the dataset that was not one of these known AR mutations will be termed a "new" SNP. As a benchmark method, each isolate was classified as being resistant or susceptible to each drug based upon the presence of any of the known AR mutations in their genome. This method was termed the "Direct Association" (DA) method, since the presence of any single one of the known AR mutations was assumed to constitute resistance.

D. Investigation of misclassified isolates

Isolates that were classified as susceptible by the Direct Association method despite being phenotypically resistant (i.e., they were false negatives because they had none of the known AR mutations) were identified. Since the resistance of these false negative isolates could not be explained by the known AR mutations, the predictive capability of any new SNPs found to be in common amongst this group was investigated. "Promising" new mutations were defined as those (a) found in at least two isolates resistant to the given drug and (b) having a positive predictive value (PPV) of 1.0.

E. Machine learning prediction

To assemble a balanced dataset for training a classifier, such that both "resistant" and "susceptible" classes were equally represented (to avoid bias), a random subset of susceptible isolates (equal to the number of resistant isolates) was selected. These resistant and susceptible isolates were randomly split 80%:20% into training and "held-out" test sets. Five-fold cross-validation within the training set was conducted to optimise the parameters of each classifier in terms of predictive accuracy. These optimised parameters were then used to train a final algorithm using all the training data (i.e., all of the 80% of the data that were previously used for cross-validation). This final algorithm was then used for prediction on the "held-out" 20% data in the test set. This process was repeated $N = 100$ times; i.e., random samplings from the pool of susceptibility examples was performed 100 times, with cross-validation and evaluation of the resulting model on the held-out test data within each iteration.

To provide a fair comparison, the performance of the Direct Association method was also assessed on the test set for each of the N iterations. The mean accuracy, sensitivity, and specificity of each method was subsequently calculated across the N iterations, allowing an assessment of the variation in the process due to the stochastic selection of the training and test data.

Prediction was performed with three different feature sets. The first consisted of all the SNPs in the dataset. The second consisted of new SNPs not known to confer resistance against any drug, and the third consisted of solely the known AR mutations for the given drug. An isolate's pattern of SNPs was thus converted into a set of features by the presence (> 0) or absence ($= 0$) of a given SNP. If a SNP was present, then a score was calculated based upon the proportion of reads that corresponded to the polymorphic base call. For instance, if 79 reads corresponded to cytosine and 25 corresponded to guanine (yet the reference base was adenine, meaning this was a SNP), then the SNP score would be $79/(79+25) = 0.76$. All analysis was conducted in python.

1) Algorithms: Logistic regression (LR) and support vector machine (SVM) classification algorithms were examined, as well as a combination of these algorithms with the Direct Association method. Detailed explanations of LR and SVM may be found elsewhere [7]. Briefly, LR is a linear classification method that optimises a set of weights $\mathbf{w}$ assigned to each input feature to provide the best classification performance among the training dataset. Its level of regularisation is governed by the parameter λ. An SVM attempts to find a separating hyperplane between two sets of labelled data points. This hyperplane's location is determined by maximising the distance between it and the closest training data points from each class, which are termed the support vectors. The adjustable parameter C determines

TABLE I

"PROMISING" NEW MUTATIONS

Drug	Mutation	N	Drug	Mutation	N
INH	gyrA_*-34*	9	RIF	rpsL_*43*	9††
	ethA_*345*	3		rpoB_insertion1299*	3
	gyrA_*456*	3	PZA	rpoB_*480*	2
	fabG1_*-17*	2		pncA_*-12*	2
	inhA_*194*	2†		pncA_*-4*	2
	rrs_*517*	2		inhA_*194*	2†

N = number of resistant isolates in the sample with the given mutation. † This is the same mutation, found for both drug searches. †† This mutation is known to cause resistance to streptomycin, another first-line MTB drug.

how heavily misclassified examples are penalised. Through the kernel trick, an SVM can be used to project data into a higher dimensional space, in which it may be linearly separable. Here, both a linear kernel and a Gaussian radial basis function (RBF) kernel were examined.

Receiver-operator curves (ROC) were constructed for both classifiers. The regularisation parameter λ (0.01 to 10) and the threshold cut-off T (0.1 to 0.9) were varied for LR classification. The cost parameter C (0.0001 to 10.0) and the radial basis function width parameter σ (0.001 to 1) were varied for SVM classification.

2) Combination prediction: To create a very sensitive classifier, a procedure was performed in which a sample was classified as being "resistant" for any case in which the Direct Association method predicted susceptibility, but for which the machine learning classifier, trained only on the "new" mutations, predicted resistance. This was conducted for both the SVM and the LR algorithms.

III. RESULTS

There were 601 unique candidate SNPs in the dataset, which occurred at 561 unique positions within the genome. Subtracting the 37 known AR mutations found in the dataset, this left 429 "new" SNPs.

A. Overall Direct Association results

Among the 650 isolates tested for INH resistance, 18 were classified as susceptible based upon the Direct Association method, despite being determined to be resistant via the phenotype tests. Similarly, 14 isolates out of the 642 tested for PZA resistance were false negatives, according to the Direct Association method.

B. Investigation of misclassified isolates

Across all drugs, 11 SNPs fit the definition of "promising" new mutations. As described in Section II-D, these were identified from the pool of new SNPs found within isolates that were phenotypically resistant, but which were lacking any of the known AR mutations. These promising mutations are shown in Table I.

C. Comparison of the performance of prediction methods

Fig. 1 shows the results of ROC analysis for a comparison between the Direct Association method, SVM, and LR classification across a range of parameters, when all SNPs (known AR and new) were included. As is evident, the Direct Association method attains a very high level of specificity. This is likely due to the large number of SNPs under consideration. Indeed, when only the set of known AR mutations are considered, the machine learning methods attain 100% specificity. However, even when all SNPs in the dataset are included, the machine learning methods are capable of superior sensitivity with only a small trade-off in specificity in comparison to the Direct Association method. In the case of INH, for instance, the Direct Association method achieves 100% specificity with 89% sensitivity; SVM prediction achieves a slightly lower 95% specificity, but with 97% sensitivity. Trading off specificity for sensitivity in this way is desirable because false negative errors are more serious than false positive ones (the optimal trade-off would also depend upon the prevalence of resistance in the given population). The performance for the drug EMB is an exception: here, the Direct Association method consistently outperforms the machine learning classifiers.

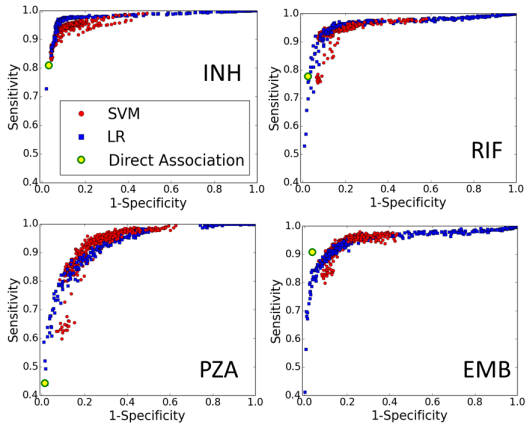

Fig. 1. Receiver-operator curves for susceptibility prediction. Sensitivity and specificity were attained by varying the parameters of the RBF SVM (C and σ) and LR (T and λ) and taking the mean across $N = 100$ subsamplings of the susceptible dataset. The mean performance of the Direct Association method is shown with a yellow circle; this method produces just a single point since it does not have any adjustable parameters.

The machine learning approach performs especially well for the drug PZA, with higher overall accuracy than the Direct Association method. This could suggest that there is an unexplained genetic contribution to resistance that was not identified by looking merely for known PZA resistance mutations. However, this could also be partly because PZA is difficult to phenotype, meaning that the gold standard phenotypic labels may be incorrect. If this is the case, then the classification algorithms may be learning an erroneous association between the mutation pattern and the phenotype.

The mean performance for each of the optimised prediction methods across $N = 100$ subsamples of the data is presented in Table II. The distribution in accuracy performance across the N subsamples is shown for INH in Fig 2. The classifier constructed by combining the Direct Association predictions with machine learning predictions (the "Comb" columns in Table II, as explained in Section II-E.2) was

TABLE II

SUMMARY OF PREDICTION PERFORMANCE ACROSS VARIOUS METHODS

		DA	SVM	SVM New	SVM Comb	LR	LR New	LR Comb
INH	Acc	**0.94**	0.93	0.88	**0.94**	0.93	0.87	0.91
	Sens	0.89	0.93	0.84	0.95	**0.98**	0.94	**0.98**
	Spec	**1.00**	0.93	0.91	0.94	0.87	0.81	0.84
RIF	Acc	**0.95**	0.84	0.79	0.88	0.85	0.79	0.84
	Sens	0.91	0.89	0.84	0.84	**1.00**	1.00	**1.00**
	Spec	**0.99**	0.80	0.74	0.77	0.70	0.59	0.67
PZA	Acc	0.78	0.82	0.72	0.79	0.82	0.75	**0.83**
	Sens	0.56	0.84	0.71	0.81	0.96	0.90	**0.98**
	Spec	**0.99**	0.80	0.74	0.78	0.69	0.61	0.68
EMB	Acc	**0.95**	0.84	0.75	0.86	0.82	0.74	0.79
	Sens	0.96	0.87	0.80	**1.00**	0.99	1.00	**1.00**
	Spec	**0.95**	0.80	0.69	0.72	0.66	0.54	0.58

Means across 100 subsamples of the data are presented. DA = Direct Association prediction method; New = classification when using only new mutations as features; Comb = combination of DA method with SVM or LR prediction. Best-performing result for each row shown in bold.

found, as expected, to be very sensitive. Perhaps surprising is the fact that the machine learning classifiers were able to attain reasonable predictive performance even when all of the known AR mutations were removed from the dataset (the "New" columns in Table II). This suggests that there is additional information contained across the remaining SNPs that is predictive of susceptibility. This could be because of shared phylogenies across susceptible and resistant strains, but it could also point towards the presence of possible compensatory or additional resistance-conferring patterns.

In terms of computational resources, the time requirement for any single drug was at most on the order of hours when run on a desktop computer, which is promising for further scaling of this approach.

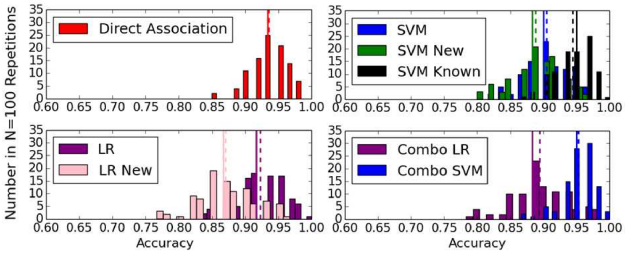

Fig. 2. Accuracy of classification methods for prediction of INH susceptibility. New = prediction using just the new SNPs; known = prediction using just the known AR mutations; combo = combination of machine learning methods with Direct Association method. The vertical solid and dashed lines correspond to the median and mean across the $N = 100$ subsamples, respectively.

IV. CONCLUSIONS

The machine learning algorithms employed here showed remarkably strong classification performance, given that they predicted bacterial susceptibility based solely upon learning the relationship between SNP patterns (in a pool of over 500 SNPs) and phenotypic susceptibility. The good classification performance, even when including as features only new SNPs (those not already known to cause resistance), also suggests that there is additional predictive information contained in the SNP pattern that is not captured by simply screening for known AR mutations.

While providing a strong proof-of-concept performance, the machine learning framework adopted here can further be improved. For instance, there is uncertainty associated with the steps of base-calling and sequence alignment. This can be incorporated into machine learning classifiers in a probabilistic manner. New, resistant strains are encountered frequently in the clinic; an online machine learning approach that could incorporate new discordant isolates would be especially desirable. The growing usage of WGS from geographically diverse locations will provide the power to look across wider regions of the genome and to identify new resistance-causing mutations in the future.

ACKNOWLEDGMENT

K.E.N. gratefully acknowledges funding from both the Rhodes Trust and the RCUK Digital Economy Programme grant number EP/G036861/1 (Centre for Doctoral Training in Healthcare Innovation). D.A.C. was funded by a Royal Academy of Engineering Research Fellowship, the Balliol Interdisciplinary Institute, and the Maurice Lubbock Memorial Fund. T.M.W. is supported by the Medical Research Council as a Clinical Training Fellow. The MTB sequences for this work were obtained through funding from the UKCRC (Wellcome Trust and MRC) and NIHR Oxford Biomedical Centre. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the RCUK, UKCRC, or NIHR.

REFERENCES

[1] X. Didelot, R. Bowden, D. J. Wilson, T. E. Peto, and D. W. Crook, "Transforming clinical microbiology with bacterial genome sequencing," *Nature Reviews Genetics*, vol. 13, no. 9, pp. 601–612, 2012.

[2] World Health Organization. (2013) Tuberculosis: Fact sheet. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs104/en/index.html

[3] C. C. Boehme, S. Saacks, and R. J. O'Brien, "The changing landscape of diagnostic services for tuberculosis," *Seminars in respiratory and critical care medicine*, vol. 34, no. 01, pp. 017–031, 2013.

[4] C. U. Köser, M. J. Ellington, E. J. Cartwright, S. H. Gillespie, N. M. Brown, M. Farrington, M. T. Holden, G. Dougan, S. D. Bentley, J. Parkhill *et al.*, "Routine use of microbial whole genome sequencing in diagnostic and public health microbiology," *PLoS pathogens*, vol. 8, no. 8, p. e1002824, 2012.

[5] T. M. Walker, C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook *et al.*, "Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study," *The Lancet infectious diseases*, 2013.

[6] M. H. Hazbón, M. Brimacombe, M. B. del Valle, M. Cavatore, M. I. Guerrero, M. Varma-Basil, H. Billman-Jacobe, C. Lavender, J. Fyfe, L. García-García *et al.*, "Population genetics study of isoniazid resistance mutations and evolution of multidrug-resistant mycobacterium tuberculosis," *Antimicrobial agents and chemotherapy*, vol. 50, no. 8, pp. 2640–2649, 2006.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.