Search for Genomic Mutations Associated with Drug-resistant Tuberculosis

Katsiaryna Rumiantsava
DMA FAMCS of
Belarussian State University
Minsk, Belarus
kattytsurikova@gmail.com

Andrei Gabrielian
Office of Cyber Infrastructure &
Computational Biology
National Institute of Allergy and
Infectious Disease, National
Institutes of Health
Bethesda, MD, USA
gabr@niaid.nih.gov

Alex Rothenthal
Office of Cyber Infrastructure &
Computational Biology
National Institute of Allergy and
Infectious Disease, National
Institutes of Health
Bethesda, MD, USA
alexr@niaid.nih.gov

Roman Sergeev
United Institute of Informatics
Problems of NAS of Belarus
Minsk, Belarus
roma.sergeev@gmail.com

Alexander Tuzikov
United Institute of Informatics
Problems of NAS of Belarus
Minsk, Belarus
tuzikov@newman.bas-net.by

Abstract. **The problem of drug-resistant tuberculosis, its diagnosis and treatment, is especially relevant today. Every year the causative agent of the disease becomes more and more resistant to existing drugs. Here we analyzed 1244 tuberculosis cases with available results of phenotypical assays for drug resistance as well as tuberculosis genome sequences using single-marker and multi-marker tests.**

Keywords: **drug-resistance, tuberculosis, phylogenetic tree, single-marker tests, multi-marker tests**

I. INTRODUCTION

The emergence of high-throughput sequencing methods for determining the DNA nucleotide sequences of living organisms has become a driving force for biological research. However, the genetic code itself does not have great practical value until the necessary information is extracted from it. The analysis of decoded genomic sequences often leads to large-scale problems, where the number of unknown parameters is measured in tens of thousands with a relatively small number of available observations. One of these tasks is the search for mutations in the genomes of microorganisms of bacterial nature, which are associated with the presence of drug resistance.

Mathematical analysis of genome-wide data on Mycobacterium tuberculosis allows predicting resistance to first-line drugs with a high probability. At the same time, resistance to second-line drugs is poorly explained only by genomic mutations. This requires a deeper study of all available data. In particular, a comparative analysis of genomic data will provide relevant information with already known cases.

II. METHODS

A. Strain collection and phylogeny

In this paper we used data collected from the Drug Resistant Tuberculosis Project [9], https://tbportals. niaid.nih.gov. At the first stage, duplicates and cases with conflicting results for drug resistance were removed and 944 cases were further analyzed. We investigated resistance to first-line drugs: isoniazid, rifampicin, pyrazinamide, ethambutol and streptomycin, as well as second-line drugs: fluoroquinolones and aminoglycosides. For each drug, we formed the two case-control group.

Case-control studies are sensitive to population separability of samples [1]. To identify population subgroups phylogenetic trees were constructed with two different methods: neighbour joining and maximum likelihood. As a result, initial data were divided into two subgroups.

To reduce the dimension of tasks and improve quality we used minor allele frequency (MAF) equal to 0.01.

B. Single-marker tests

Single-marker tests are used to test associations between observed drug resistance and individual mutations [2]. We used modifications of classical statistical tests as single-marker tests: Fisher's exact test and Cochran-Mantel-Haenszel test. Both methods are based on building contingency tables of the following form.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

TABLE I. CONTINGENCY TABLE CONSIDERED IN SINGLE-MARKER TESTS TO SEARCH FOR MUTATIONS ASSOCIATED WITH DRUG RESISTANCE

drug susceptibility	Presence of mutation		
	present	*absent*	*total*
sensitive	n_{00}	n_{01}	n_{0*}
resistant	n_{10}	n_{11}	n_{1*}
total	n_{*0}	n_{*1}	n_{**}

The Cochran-Mantel-Haenszel test, in contrast to the Fisher test, takes into account population subgroups. We need to make adjustments for the population subdivision because the variation in the frequency of occurrence of some mutations can be explained by their belonging to different populations not by drug susceptibility. To reduce the likelihood of errors in multiple hypothesis testing, we used the Bonferroni correction.

C. Multi-marker tests

Multi-marker tests unlike single-marker consider additive effects between mutations. In this study, we used an algorithm for searching for combinations of mutations boosting [1] and factor analysis of mixed data with linear mixed model.

In addition to genomic data, we also considered phenotypic traits of samples. Such data include both categorical and continuous features. Therefore, we used factor analysis of mixed data (FAMD) to transform them into a set of uncorrelated features.

The linear mixed model is used for regression on hierarchical data [4]. It considers population subgroups in original data. The main idea of the linear mixed model is that it takes into account both fixed and random effects. The fixed effects are the basic regression on the data. Each subgroup may have its own unique characteristics or traits, which can be expressed in presence of common mutations that are not associated with drug resistance, and vice versa, the peculiarities of diagnosis and treatment of tuberculosis and the health system itself. There may also be similarities in the closest subgroups. The linear mixed model considers such subgroup effects as random effects. The model can be described with the equation:

$$Y = X\beta + Z\upsilon + \varepsilon. \qquad (1)$$

Where Y is a phenotype vector, X – matrix with data, β – vector of effects that determine the significance of a set of mutations, Z – matrix of random effects of subgroups, υ – vector of random effects for subgroups, ε – vector of residues that cannot be explained by the model.

To control how well the model works we used cross-validation for a given number of blocks [5].

Another approach we used in the study is the search for combinations of mutations. The main feature of this task is that total number of considered mutations significantly exceeds the number of observations. The task of enumerating all the possible combinations of mutations requires large computational resources. To avoid this problem, one can switch from the task of enumerating mutations to the task of enumerating samples, since their number is much smaller. An exhaustive description of the algorithm is given in [1, 3].

III. RESULTS

First, we carried out analysis of the results of biological tests for drug resistance that revealed cross-resistance between drugs. Selected correlation coefficients calculated from the results of biological tests are presented in Fig. 1.

	isoniazid	rifampicin	streptomycin	ethambutol	pyrazinamide	ofloxacin	levofloxacin	capreomycin	amikacin	kanamycin
isoniazid	1.000000	0.892679	0.784130	0.571943	0.293037	0.207332	0.107483	0.091817	0.077532	0.250390
rifampicin	0.892679	1.000000	0.706333	0.552775	0.281498	0.211742	0.136717	0.105422	0.088985	0.244520
streptomycin	0.784130	0.706333	1.000000	0.515814	0.234472	0.168137	0.149634	0.152227	0.102939	0.212197
ethambutol	0.571943	0.552775	0.515814	1.000000	0.503125	0.286219	0.297925	0.262457	0.308009	0.446664
pyrazinamide	0.293037	0.281498	0.234472	0.503125	1.000000	0.310293	0.445298	0.222662	0.261485	0.271807
ofloxacin	0.207332	0.211742	0.168137	0.286219	0.310293	1.000000	0.733575	0.421209	0.357054	0.394190
levofloxacin	0.107483	0.136717	0.149634	0.297925	0.445298	0.733575	1.000000	0.650301	0.495414	0.646352
capreomycin	0.091817	0.105422	0.152227	0.262457	0.222662	0.421209	0.650301	1.000000	0.674982	0.595912
amikacin	0.077532	0.088985	0.102939	0.308009	0.261485	0.357054	0.495414	0.674982	1.000000	0.709660
kanamycin	0.250390	0.244520	0.212197	0.446664	0.271807	0.394190	0.646352	0.595912	0.709660	1.000000

Fig. 1. Pairwise correlations between biological test results

In this work, we constructed two phylogenetic trees for 46 genes of 944 sequences using different approaches: neighbour joining [6] and maximum likelihood [7]. Despite the difference in the lengths of branches and variation in the locations of the end nodes, identical subgroups are identified in both trees.

We used the R software functions from the stats package: fisher_test for Fisher's exact test and mantelhaen_test for the Cochran-Mantel-Haenszel test. We calculated p-values for all corresponding mutations in the data matrix and consider only those with p-value greater than 5×10^{-8}. To visualize test results Manhattan charts were used. The diagram displays a set of points, the coordinate of which on the abscissa is the position of the mutation, and on the ordinate is the negative logarithm of the p-value. Most of the points on the graph are close to abscissa axis. The higher points along the ordinate axis, the lower are their p-values, which mean the greater statistical significance.

To validate the results obtained, we compared them with the drug resistance mutations published in [8].

Single-marker tests revealed a set of mutations associated with drug resistance to individual first-line drugs, their combinations, as well as groups of second-line drugs. However, they were unable to identify

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

mutations associated with resistance to individual second-line drugs, but only their combinations.

For factor analysis of mixed data, we used the prince python package, the implementation of a mixed linear model from the statsmodel package; cross-validation of the generalizability of the model was performed using StratifiedKFold over five partitions from the sklearn package. To validate classification results we used four classical metrics in machine learning: precision, recall, f-score and accuracy.

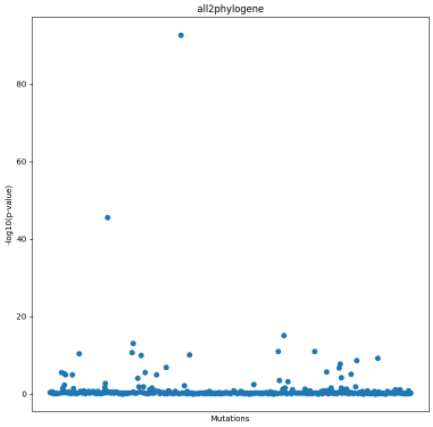

Fig. 2. Manhattan Plot of Cochran-Mantel-Haenszel test results (example) for first-line drugs

For first-line drugs and their combinations, except ethambutol, all metrics values were greater than 0.8. The percentage of true positives among all positives is high (95% and above) for all such samples. The proportion of predicted drug resistance among all is about 80-90%. The proportion of correctly classified samples is slightly higher than the previous one. It is important that the classifier covers as many cases of drug resistance as possible and only few sensitive.

For the second-line drugs, more or less good metric values were obtained for fluoroquinolones, including ofloxacin. The accuracy of the classification is about 80%; the recall is a bit lower, about 75%. For aminoglycosides, the classification results are much worse. For capreomecin the algorithm failed to build the model. For amikacin, the percentage of true positives among all positives turned out to be below 50%, which is equivalent to random labeling. The classification results are slightly better for kanamycin, on average, all metrics values were in the region of 0.66 - 0.71, but these are also low estimates. Based on the classification results, it can be concluded that it was not possible to build a model for aminoglycosides that could predict drug resistance to these drugs with good probability.

In the study, we used H37Rv NC_000962.3 reference sequence; all mutations mentioned below correspond to that sequence. For all first-line drugs and their combinations, the mutation combination search algorithm found one dominant mutation, C2155175G, associated with isoniazid resistance. When boosting on these sets, the algorithm added to the dominant mutations new ones related to the markers of phylogenetic lines, while the accuracy of the classifiers did not increase.

Considering all mutations occurring in one position as one mutation, the algorithm has identified a set of mutations: 7570, 7572, 7581, 7582 associated with resistance to fluoroquinolones (according to information from TBDreamDB [8]). Despite the low accuracy, the result for fluoroquinolones, except levofloxacin, is considered satisfactory.

For aminoglycosides, with the exception of kanamycin, scores were rather low. Each set of mutations contains A1473252 associated with resistance to this group of drugs. This is consistent with the results of single-marker tests. However, this mutation, even when combined with others, is not enough to build a good classifier. For kanamycin, the algorithm constructed a set of mutations, which included the G2715356A mutation associated with resistance to this particular drug.

IV. Discussion

In this paper, we considered the problem of drug-resistant tuberculosis and solutions for the comparative analysis of mycobacterial genomes. Using the phylogenetic tree, two population subgroups were identified in the original dataset. We found cross-resistance between the individual drugs.

To search for mutations associated with drug resistance, we used single-marker and multi-marker tests. To validate the test results, the TBDreamDB database was used, which contains already known mutations associated with drug resistance. Single-marker tests revealed a set of mutations associated with resistance to individual first-line drugs, their combinations, as well as groups of second-line drugs. However, they were unable to identify mutations associated with resistance to individual second-line drugs. Multi-marker tests built good classifiers and identified combinations of mutations for individual first-line drugs and their groups. For second-line drugs, multi-marker tests have built combinations that include individual mutations associated with drug resistance, but they alone are not enough to build a good classifier.

Acknowledgment

We acknowledge support from TB Portals Consortium and the TB Portals Program.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

References

[1] R. Sergeev, Algorithms for analysis and search for associations in genetic data, PhD thesis, Minsk, 2019.

[2] X. Wang, N. J. Morris, D. J. Schaid, R. C. Elston, Power of single- vs. multi-marker tests of association, Genet Epidemiol, 2012, 36 (5), pp. 480-487.

[3] K. Tsurykava, R. Sergeev, Algorithm for searching combinations of mutations associated with drug-resistant tuberculosis, Minsk, BSU, 2021, pp. 130-132.

[4] J. C. Pinheiro and D. M. Bates, Mixed-effects models in S and S-PLUS, New York, NY u.a., Springer, 2000.

[5] J. Fox, Applied Regression Analysis and Generalized Linear Models. Sage Publications, Thousand Oaks, California, 2015.

[6] Rapid Neighbour Joining, Martin Simonsen, Thomas Mailund and Christian N. S. Pedersen, In Proceedings of the 8th Workshop in Algorithms in Bioinformatics (WABI), LNBI 5251, Springer Verlag, October 2008, pp. 113-122.

[7] M. Salemi, A.-M. Vandamme, Lemey, The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing, Cambridge, UK, Cambridge University Press, 2009.

[8] A. Sandgren, M. Strong, P. Muthukrishnan, Weiner BK, Church GM, Murray MB (2009) Tuberculosis Drug Resistance Mutation Database. PLoS Med 6(2): e1000002. https://doi.org/10.1371/journal.pmed.1000002.

[9] Alex Rosenthal, Andrei Gabrielian, Eric Engle, et. al. The TB Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis. Journal of Clinical Microbiology, vol. 77, no. 1, 2017, pp. 3261-3282. https://doi.org/10.1128/JCM.01013-17.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.