A NOVEL ALGORITHM FOR PREDICTION OF DRUG RESISTANCE USING CROSS-ACCURACY COMPUTATION

Yu-Xiang Chen[1], Alexander V. Tuzikov[2]

1) Belarusian State University, Minsk, Belarus
2) United Institute of Informatics problems of the National Academy of Sciences, Minsk, Belarus
E-mail: 1c894424323@outlook.com, 2tuzikov@newman.bas-net.by

Genome-wide association studies (GWAS) are vital in determining disease etiology. Since most of the existing computational models for GWAS require many features, they are often too computationally expensive for clinical use. An algorithm for prediction of *Mycobacterium tuberculosis* drug resistance is proposed by using cross-accuracy computations. We do exhaustive investigation of SNP pairs followed by a greedy algorithm to improve prediction accuracy of SNP combinations. A dataset of 3178 *Mtb* genomes was tested to predict resistance to 8 drugs. The results were compared with those from TB-profiler and Mykrobe web systems. They indicate that our algorithm closely approximates the performance of the existing methods for first-line drugs, while is outperforming them in accuracy for the second-line drugs tested.

Keywords: drug-resistant tuberculosis; SNPs combination; cross-accuracy.

НОВЫЙ АЛГОРИТМ ПРЕДСКАЗАНИЯ ЛЕКАРСТВЕННОЙ УСТОЙЧИВОСТИ НА ОСНОВЕ ВЫЧИСЛЕНИЯ ТОЧНОСТИ ДЛЯ ПАР МУТАЦИЙ

Юйсян Чэнь[1], Александр Тузиков[2]

[1]Белорусский государственный университет, Минск, Беларусь
[2] Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь
[1]c894424323@outlook.com, [2]tuzikov@newman.bas-net.by

Полногеномный анализ ассоциаций (ПАА) имеет важное значение для определения этиологии заболеваний. Поскольку большинство существующих вычислительных моделей для ПАА требуют вычисления многих признаков, они зачастую слишком трудоемкие для использования в клинической практике. В докладе предложен алгоритм прогнозирования лекарственной устойчивости микобактерий туберкулеза (*Mtb*), основанный на вычислении точности предсказания для всех пар мутаций. Для повышения точности предсказания на основе комбинаций мутаций используется жадный алгоритм. Набор данных из 3178 геномов *Mtb* был протестирован для предсказания устойчивости к 8 лекарствам. Полученные результаты сравнивались с результатами веб-сервисов TB-profiler и Mykrobe. Они показывают, что предложенный алгоритм незначительно уступает по точности предсказания этим сервисам для препаратов первой линии, но превосходит их по точности для протестированных препаратов второй линии.

Ключевые слова: лекарственно-устойчивый туберкулез; комбинация мутаций; перекрестная точность.

INTRODUCTION

Genome-wide association analysis (GWAS) is an algorithm that searches for correlations between single nucleotide polymorphisms (SNPs) and traits [1]. Due to the large number of SNPs, finding combinations of k SNPs among hundreds of thousands in genomes is a complex combinatorial problem. In recent years, Monte Carlo method [2], spanning tree method [3], machine learning methods [4, 5] etc., have appeared, which can improve search efficiency. Discriminant functions for calculating associations between SNP combinations and phenotypes are essential, and discriminant functions with light computations are favorable for improving search efficiency. Mutual information and conditional entropy (Shannon entropy-based methods) [6] and chi-square tests [7] are widely used to assess associations in a lightweight way.

The aim of this research is to explore a search method that is able to find SNP combinations satisfying high recognition accuracy. The methodology developed uses exhaustive investigation of SNPs interaction for all SNP pairs followed by a greedy algorithm to find combinations of SNPs associated with phenotypes.

METHODS

If an initial sequence of SNPs $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ contains thousands of elements, then one can compute accuracy values $a(x_i, x_j)$ (1) for combined SNPs created by all pairs (x_i, x_j) of single SNPs. It allows to use further this information for finding the most important combinations of SNPs.

Let us select q pairs (x_i, x_j) with the highest $a(x_i, x_j)$ values which contain distinct SNPs. Now we will propagate these pairs of SNPs to combinations of up to l SNPs by a greedy algorithm as follows. For every selected pair (x_i, x_j) consider combinations of three SNPs (x_i, x_j, x_r) for every x_r from the initial sequence $\mathbf{x}$ distinct from x_i and x_j. Select an arbitrary combination of three SNPs (x_i, x_j, x_k) with the maximum accuracy value $a(x_i, x_j, x_k)$. Similarly, one can get combinations of up to l SNPs starting from the selected pair (x_i, x_j). Finally, let us select either a single SNP or a combination of SNPs associated with resistance to a considered drug with the maximum accuracy value.

The quality of prediction of a phenotype based on a SNP in a considered genome position can be evaluated using several measures:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}, \text{Specificity} = \frac{TN}{TN + FP}, \\ \text{Precision} = \frac{TP}{TP + FP}, \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Prediction of drug resistance is based on the presence of a SNP under consideration in a genome sample. If a sample is resistant to drug d and is predicted correctly, then the prediction is considered true positive (TP). Similarly, if a sample is sensitive to drug d and is predicted to be susceptible, then the prediction is considered true negative (TN). However, no prediction is perfect, and if a sample is resistant but is predicted to be sensitive, then the prediction is considered false negative (FN). Similarly, if a sample is sensitive do drug d but is predicted to be resistant, then the prediction result is false positive (FP).

RESULT AND DISCUSSION

The original data set used in this study included the drug susceptibility test data (DST) and genome-wide data corresponding to these cases. These data were taken from the TB portal [8], presenting an excellent platform for investigations on drug-resistant tuberculosis (TB). The DST data provide verified information on resistance or sensitivity of *Mtb* samples to considered drugs. Our dataset contained 3178 whole genomes. We investigated resistance to 4 first-line drugs such as Isoniazid (INH), Rifampin (RIF), Streptomycin (SM), Ethambutol (EMB), and 4 second-line drugs, such as Amikacin (AMK), Linezolid (LZD), Levofloxacin (LFX) and Kanamycin (KM).

The efficiency of the developed algorithm was compared with Mykrobe [9] and TB-Profiler [10]. Both are considered as practical prediction tools for *Mycobacterium tuberculosis* drug resistance. The prediction results of "Mykrobe v0.13.0" and "TB-Profiler v5.0.1" were computed for the same dataset.

Our algorithms showed slightly underperformance than Mykrobe and TB-profiler for first-line drugs, however, for second-line drugs we have got higher accuracy values than Mykrobe and TB-Profiler. The algorithm is clearly applicable to maximize other predictors as well. It has also an advantage of providing results that are clearly interpretable, since resistance to a specific drug is predicted by the presence of the corresponding individual mutations in a genome sample under consideration.

Comparison of the developed algorithms with Mykrobe and TB-profiler systems

Drug	Recall	Specificity	Precision	Accuracy	Method
INH	90.3%	91.8%	96.1%	90.7%	Our algorithm
	89.4%	95.7%	97.9%	91.3%	Mykrobe
	89.9%	95.2%	97.7%	91.5%	TB-profiler
RIF	93.8%	80.9%	87.9%	88.6%	Our algorithm
	92.6%	84.8%	90.1%	89.5%	Mykrobe
	93.8%	84.5%	90.0%	90.1%	TB-profiler
SM	84.8%	69.9%	82.1%	79.1%	Our algorithm
	94.7%	64.7%	81.4%	83.3%	Mykrobe
	93.0%	66.2%	81.7%	82.8%	TB-profiler
EMB	91.9%	70.3%	69.5%	79.5%	Our algorithm
	86.5%	78.8%	75.0%	82.1%	Mykrobe
	93.2%	72.0%	71.0%	81.0%	TB-profiler
AMK	94.9%	58.3%	75.9%	79.5%	Our algorithm
	32.6%	95.3%	90.6%	58.9%	Mykrobe
	33.2%	95.0%	90.2%	59.1%	TB-profiler
LFX	92.7%	49.2%	73.6%	75.5%	Our algorithm
	50.5%	90.0%	88.5%	66.1%	Mykrobe
	53.5%	89.4%	88.6%	67.7%	TB-profiler
LZD	93.0%	80.1%	83.3%	86.7%	Our algorithm
	0.3%	100.0%	100.0%	48.6%	Mykrobe
	0.4%	100.0%	100.0%	48.6%	TB-profiler
KM	70.0%	80.0%	75.7%	75.3%	Our algorithm
	78.1%	68.1%	68.5%	72.8%	Mykrobe
	79.8%	66.6%	68.0%	72.8%	TB-profiler

CONCLUSION

We proposed a novel algorithm for predicting *Mycobacterium tuberculosis* drug resistance using cross-accuracy computations. The results obtained demonstrate the superiority of the developed algorithm over Mykrobe and TB-Profiler for the second-line drugs tested. They showed also that the number of SNPs in combinations can be constrained by 5 to ensure an appropriate approximation of prediction accuracy using the developed algorithm. This can also be utilized to reduce the computational efforts required to search for SNP combinations.

REFERENCES

1. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. / Hindorff L.A. pet al.] // PNAS. 2009. P. 9362-9367. DOI: 10.1073/pnas.0903103106

2. A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. / Wang J. [et al.] // BMC Genomics. 2015. Vol. 16. P. 1-20. DOI: 10.1186/s12864-015-2217-6.
3. TEAM: efficient two-locus epistasis tests in human genome-wide association study. / Zhang X. [et al.] // Bioinformatics. 2010. P. i217-i227. DOI: 10.1093/bioinformatics/btq186.
4. TSLRF: two-stage algorithm based on least angle regression and random forest in genome-wide association studies. / Sun J. [et al.] // Scientific reports. 2019. Vol. 9. P. 18034. DOI: 10.1038/s41598-019-54519-x
5. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. / Mieth B. [et al.] // Scientific reports. 2016. Vol. 6. P. 36671. DOI: 10.1038/srep36671
6. ELSSI: parallel SNP–SNP interactions detection by ensemble multi-type detectors. / Wang X. [et al.] // Briefings in Bioinformatics. 2022. Vol. 23. P. bbac213. DOI: 10.1093/bib/bbac213.
7. Distributed multi-objective optimization for SNP-SNP interaction detection. / Li F. [et al.] // Methods. 2024. Vol. 221. P. 55-64. DOI: 10.1016/j.ymeth.2023.11.016
8. The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis. / Rosenthal A. [et al.] // Journal of clinical microbiology. 2017. Vol. 55. P. 3267-3282. DOI: 10.1128/jcm.01013-17
9. Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe. / Hunt M. [et al.] // Wellcome open research. 2019. Vol. 4. P. 191. DOI: 10.12688/wellcomeopenres.15603.1
10. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. / Phelan J. E. [et al.] // Genome medicine. 2019. Vol. 11. P. 41. DOI: 10.1186/s13073-019-0650-x