
ДИСКРЕТНАЯ МАТЕМАТИКА И МАТЕМАТИЧЕСКАЯ КИБЕРНЕТИКА

DISCRETE MATHEMATICS AND MATHEMATICAL CYBERNETICS

УДК 519.168, 575.174

ИДЕНТИФИКАЦИЯ КОМБИНАЦИЙ ГЕНОМНЫХ МУТАЦИЙ С ПОМОЩЬЮ ПОЛНОГЕНОМНОГО ПОИСКА АССОЦИАЦИЙ НА ПРИМЕРЕ МИКОБАКТЕРИИ ТУБЕРКУЛЕЗА

ЧЭНЬ ЮЙСЯН¹⁾, А. М. АНДРИАНОВ²⁾, А. В. ТУЗИКОВ¹⁾

¹⁾Объединенный институт проблем информатики НАН Беларуси,
ул. Сурганова, 6, 220012, г. Минск, Беларусь

²⁾Институт биоорганической химии НАН Беларуси,
ул. Академика Купревича, 5, корп. 2, 220084, г. Минск, Беларусь

Аннотация. Полногеномный поиск ассоциаций играет ключевую роль в выявлении взаимосвязей между геномами и фенотипами. Многие исследования в этой области посвящены изучению генетических вариаций и их взаимодействий в геномах. Однако, несмотря на достигнутый значительный прогресс в данном направлении, рассматриваемая проблема по-прежнему является крайне актуальной и требует разработки эффективных методов

Образец цитирования:

Чэнь Юйсян, Андрианов АМ, Тузиков АВ. Идентификация комбинаций геномных мутаций с помощью полногеномного поиска ассоциаций на примере микобактерии туберкулеза. *Журнал Белорусского государственного университета. Математика. Информатика.* 2024;3:73–89. EDN: PTGHLY

For citation:

Chen Yu-Xiang, Andrianov AM, Tuzikov AV. Identification of combinations of genomic mutations using genome-wide association studies by example of *Mycobacterium tuberculosis*. *Journal of the Belarusian State University. Mathematics and Informatics.* 2024;3:73–89. Russian. EDN: PTGHLY

Авторы:

Чэнь Юйсян – младший научный сотрудник лаборатории математической кибернетики.

Александр Михайлович Андрианов – доктор химических наук, профессор; главный научный сотрудник лаборатории белковой инженерии.

Александр Васильевич Тузиков – доктор физико-математических наук, член-корреспондент НАН Беларуси, профессор; заведующий лабораторией математической кибернетики.

Authors:

Chen Yu-Xiang, junior researcher at the laboratory of mathematical cybernetics.

c894424323@outlook.com

Alexander M. Andrianov, doctor of science (chemistry), full professor; chief researcher at the laboratory of protein engineering. alexande.andriano@yandex.ru

Alexander V. Tuzikov, doctor of science (physics and mathematics), corresponding member of the National Academy of Sciences of Belarus, full professor; head of the laboratory of mathematical cybernetics.

tuzikov@newman.bas-net.by

и алгоритмов ее решения. Для поиска ассоциированных с фенотипом комбинаций однонуклеотидных полиморфизмов в настоящей статье предложены четыре новых алгоритма, основанных на изучении взаимодействия однонуклеотидных полиморфизмов в двух режимах – аддитивном и мультипликативном. На первом этапе эти алгоритмы используют полный перебор пар однонуклеотидных полиморфизмов для предсказания их ассоциации с фенотипом, а на втором этапе – жадные процедуры для поиска комбинаций, включающих до пяти однонуклеотидных полиморфизмов с наибольшими величинами ассоциации. Разработанный вычислительный подход протестирован на наборе данных, содержащем 3178 геномов микобактерии туберкулеза, для выявления комбинаций мутаций и прогнозирования устойчивости различных штаммов микобактерии туберкулеза к 20 лекарственным препаратам. Полученные результаты сопоставлены с результатами прогнозирования лекарственной устойчивости микобактерии туберкулеза современными программными системами *Mykrobe* и *TB-Profiler*. Для 5 препаратов первой линии и 1 препарата второй линии (офлоксацина) системы *Mykrobe* и *TB-Profiler* по правильности предсказания несколько превосходят предложенные авторами алгоритмы, однако для остальных 14 препаратов второй линии уступают им.

Ключевые слова: полногеномный поиск ассоциаций; взаимодействие однонуклеотидных полиморфизмов; лекарственно-устойчивый туберкулез.

Благодарность. Работа выполнена при поддержке консорциума по лекарственно-устойчивому туберкулезу (<https://tbportals.niaid.nih.gov>), Международного научно-технического центра (проект PR150), а также Китайского совета по стипендиям (<https://www.csc.edu.cn/>).

IDENTIFICATION OF COMBINATIONS OF GENOMIC MUTATIONS USING GENOME-WIDE ASSOCIATION STUDIES BY EXAMPLE OF *MYCOBACTERIUM TUBERCULOSIS*

CHEN YU-XIANG^a, A. M. ANDRIANOV^b, A. V. TUZIKOV^a

^aUnited Institute of Informatics Problems, National Academy of Sciences of Belarus,
6 Surganova Street, Minsk 220012, Belarus

^bInstitute of Bioorganic Chemistry, National Academy of Sciences of Belarus,
5 Akademika Kuprevicha Street, 2 building, Minsk 220084, Belarus

Corresponding author: A. V. Tuzikov (tuzikov@newman.bas-net.by)

Abstract. Genome-wide association studies play a key role in identifying relationships between genomes and phenotypes. Many studies in this field are devoted to the investigations of genetic variations and their interactions in genomes. However, despite significant progress achieved in this direction, the problem under consideration is still highly relevant and requires the development of effective methods and algorithms for solving it. In this paper, four new algorithms based on the study of single nucleotide polymorphism interactions in the two modes, additive and multiplicative, are proposed to find combinations of single nucleotide polymorphisms associated with phenotypes. In the first stage, the algorithms use exhaustive search of single nucleotide polymorphism pairs to predict their association with phenotype, and in the second stage, greedy procedures are applied to find combinations of up to five single nucleotide polymorphisms with the best association values. The developed computational approach is tested on the dataset containing 3178 *Mycobacterium tuberculosis* genomes to identify single nucleotide polymorphism combinations and predict resistance of *Mycobacterium tuberculosis* strains to 20 drugs. The results obtained are compared with those of the modern prediction software systems *Mykrobe* and *TB-Profiler*. For the 5 first-line drugs and the 1 second-line drug (Ofloxacin), *Mykrobe* and *TB-Profiler* systems slightly exceed the prediction accuracy of the proposed algorithms, but for the other 14 second-line drugs, they are inferior to them.

Keywords: genome-wide association studies; GWAS; single nucleotide polymorphism interaction; drug-resistant tuberculosis.

Acknowledgements. This work was supported by the drug-resistant tuberculosis consortium (<https://tbportals.niaid.nih.gov>), the International Science and Technology Centre (project PR150), as well as the China Scholarship Council (<https://www.csc.edu.cn/>).

Введение

Полногеномный поиск ассоциаций (*genome-wide association studies*, GWAS) включает методы, используемые для выявления взаимосвязей между геномами и фенотипами. Основная цель этих исследований – понимание биологии фенотипов, расширение знаний и разработка новых алгоритмов анализа геномных данных. С помощью GWAS получено большое количество результатов в области популяционной генетики и генетики сложных признаков, биологии заболеваний, их диагностики и лечения [1].

Механизмы взаимосвязей между геномами и фенотипами сложны, и многие из них пока не имеют достоверного объяснения. Однако за последние почти 20 лет был достигнут значительный прогресс в их понимании и использовании [1; 2].

Ряд исследований основаны на изучении взаимосвязи между генетическими вариациями и фенотипами. Однонуклеотидные полиморфизмы (ОНП) в геномах могут вызывать фенотипические проявления или быть связанными с ними. Хорошо известно, что большинство фенотипов ассоциированы не с одной генетической вариацией в конкретной геномной позиции, а со многими из них. Изучение взаимодействия ОНП играет ключевую роль в GWAS. При этом проблема выявления подобных зависимостей представляет собой сложную комбинаторную задачу. Учитывая тот факт, что количество генетических вариаций в геномах может измеряться сотнями тысяч или даже миллионами, задача изучения таких зависимостей становится неразрешимой с помощью современных компьютеров. Разработанные подходы к решению данной проблемы либо исследуют все ОНП, либо выделяют из этого множества подмножества с нужными свойствами, которые затем используются для выявления маркеров с лучшими прогностическими характеристиками. Полный перебор (исчерпывающий поиск) комбинаций ОНП – простой и надежный метод, который предсказывает взаимосвязи между генетическими вариациями и фенотипами. Однако из-за огромной вычислительной сложности он не может найти широкого применения. По этой причине был разработан алгоритм с поддержкой взаимодействия ОНП на графических картах, протестированный на нескольких наборах данных [3]. В работе [4] генетический алгоритм и SVM-классификатор применяли для поиска подмножеств ОНП, наиболее предпочтительных с точки зрения классификатора, а в работе [5] был предложен алгоритм случайного леса для отбора подмножеств ОНП, связанных с фенотипами.

Один из популярных подходов к поиску ассоциаций между геномными мутациями и фенотипами заключается в объединении мутаций генов и изучении проблемы ассоциации между генами и фенотипами [6]. В этом случае размерность задачи существенно уменьшается, и появляются новые возможности для биологической интерпретации фенотипов. Данному подходу посвящено множество публикаций. В частности, в работе [7] предложен новый инструмент, использующий сводную статистику для обнаружения ассоциаций на уровне генов, а не отдельных ОНП. Этот инструмент позволяет значительно сократить объем вычислений и является более мощным, чем другие программные средства, применяющие аналогичный подход. Взвешивание генетических вариаций в гене – одна из основных проблем при разработке подходов к поиску ассоциаций на уровне генов, влияющая на их эффективность. Генный тест, также использующий сводную статистику и названный OWC, продемонстрировал более высокую мощность на модельных и реальных данных по сравнению с другими популярными методами [8]. Изучение взаимодействия генов и их совместного влияния на сложные фенотипы является актуальной темой современных исследований [9].

Микобактерия туберкулеза (МБТ) – возбудитель инфекционного заболевания, называемого туберкулезом. В 2022 г. им заболели 10,6 млн человек во всем мире, из них у 410 тыс. человек развился туберкулез с множественной лекарственной устойчивостью или устойчивый к рифампицину туберкулез, а 1,3 млн человек умерли. Число диагностированных и начавших лечение людей было намного меньше, чем предполагаемое число заболевших [10]. Методы прогнозирования лекарственной устойчивости на основе анализа ОНП могут предоставить врачам ценную информацию для рекомендации более эффективных препаратов и схем лечения, необходимых для сокращения числа случаев лекарственной устойчивости [11; 12]. В исследовании [13] взаимодействия ОНП были проанализированы для предсказания лекарственной устойчивости МБТ и выявления ассоциированных с ней пар генов. Одномерная сверточная нейронная сеть превзошла другие традиционные классификаторы машинного обучения в поиске комбинаций ОНП, связанных с лекарственной устойчивостью МБТ [14]. Согласно результатам исследования [14] для 8 препаратов эта нейронная сеть обеспечила более высокую точность прогнозирования, чем современная программа предсказания лекарственной устойчивости МБТ *Mykrobe* [15].

За последнее десятилетие было разработано много моделей для предсказания лекарственной устойчивости [15–23]. Однако задача корректного прогнозирования устойчивости к лекарствам на основе информации о наличии определенных мутаций в геноме по-прежнему остается крайне актуальной и требует разработки новых подходов к ее решению.

В настоящем исследовании для прогнозирования лекарственной устойчивости МБТ использован поиск ОНП в геномах на основе данных портала по туберкулезу (<https://tbportals.niaid.nih.gov>), который включает экспериментальную информацию об устойчивости к 20 лекарственным препаратам для 3178 геномов МБТ, содержащихся в базе данных GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>). Для этого были разработаны четыре алгоритма поиска комбинаций ОНП, ассоциированных с лекарственной устойчивостью МБТ. Данные алгоритмы основаны на изучении взаимодействия ОНП в двух режимах – аддитивном и мультипликативном. На первом этапе алгоритмы используют полный перебор пар ОНП для расчета

их связи с устойчивостью к 20 лекарственным препаратам, а на втором этапе – жадные процедуры для поиска комбинаций, включающих до пяти ОНП, которые наилучшим образом связаны с лекарственной устойчивостью.

Материалы и методы исследования

Задача GWAS. Для каждого лекарственного препарата d исследуемые образцы подразделяются на устойчивые и чувствительные к нему. Таким образом, эталонный набор данных S геномов МБТ можно представить в виде

$$S = S_d^+ \cup S_d^-,$$

где S_d^+ – подмножество образцов, устойчивых к препарату d , а S_d^- – подмножество образцов, чувствительных к препарату d .

В рассматриваемом случае каждый образец S_i , $i = 1, 2, \dots, m$, представляет собой последовательность, состоящую из четырех нуклеотидов $s_{i,j} \in \{A, T, C, G\}$:

$$S_i = (s_{i,1}, s_{i,2}, \dots, s_{i,n}).$$

Здесь n – количество нуклеотидов в геноме. Величина n отличается для разных организмов. Например, для генома МБТ $n = 4\,418\,596$. Последовательность нуклеотидов в геноме может содержать различные однопозиционные мутации, обычно называемые ОНП, означающие замену одного нуклеотида в определенной позиции на другой нуклеотид и присутствующие в достаточно большой части популяции (более 1 %). Некоторые ОНП важны для жизни организма, а также связаны с его фенотипом (или признаком), в рассматриваемом случае с устойчивостью микроорганизма к какому-либо лекарству. Предполагается, что задан вектор фенотипа $y_d = (y_1, y_2, \dots, y_m)^T$, где m – размер набора данных S , и $y_i = 1$, если образец S_i устойчив к препарату d . В противном случае $y_i = 0$ и образец S_i чувствителен к данному препарату.

Рассматриваемая в статье задача GWAS заключается в поиске в геноме ОНП, ассоциированных с фенотипами, если такие ОНП существуют. В этом случае размерность задачи может быть уменьшена до анализа только последовательности ОНП, а не всего генома. Предположим, что набор геномных данных S содержит последовательность $x = (x_1, x_2, \dots, x_p)$, состоящую из p ОНП, полученных при сравнении геномов из базы данных с эталонным геномом, причем каждый x_i , $i = 1, 2, \dots, p$, соответствует некоторому ОНП в образцах генома. Отметим, что рассматриваемые ОНП встречаются в не менее чем 1 % геномов из набора данных S . Тогда вектор $x_{i,0} = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$ описывает значения ОНП для генома S_i , $x_{i,k} = 1$, если ОНП с номером $k \leq p$ существует в этом геноме, и $x_{i,k} = 0$ в противном случае. Все ОНП в наборе данных S для препарата d определяются матрицей ОНП

$$X_d = (x_{i,j})_{m \times p}. \quad (1)$$

В статье рассматривается следующая задача: по матрице ОНП X_d (1) и вектору фенотипов y_d следует идентифицировать отдельные ОНП или их комбинации, связанные с фенотипом.

Сначала исследуем взаимосвязи между отдельными ОНП, представленными в наборе образцов и заданными векторами $x_{0,j} = (x_{1,j}, x_{2,j}, \dots, x_{m,j})^T$ для $j = 1, 2, \dots, p$, а именно между столбцами матрицы ОНП X_d (1) и вектором фенотипа y_d . Такой подход называется одномолекулярным тестом. Далее исследуем взаимосвязи между комбинациями нескольких ОНП и вектором фенотипа, т. е. выполним многомаркерный тест. Во втором случае задача приобретает комбинаторный характер, поэтому для проверки разных комбинаций ОНП будем использовать эвристики. Можно объединить вектор-столбцы, соответствующие комбинации двух ОНП, с помощью различных операций и ввести в матрицу ОНП X_d (1) объединенный ОНП. В данной работе для каждой координаты вектор-столбцов при объединении соответствующих вектор-столбцов отдельных ОНП используются логические операции «или» либо «и». Операция «или» предполагает аддитивный эффект отдельных ОНП для предсказания фенотипа, такое взаимодействие ОНП называется аддитивным. Использование операции «и» означает, что взаимодействие мультипликативно и предсказание фенотипа происходит, когда оба одиночных ОНП предсказывают этот фенотип.

Исходные данные. Набор данных S содержал 3178 геномов и результаты тестов на устойчивость к 20 лекарственным препаратам, включая 5 препаратов первой линии, а именно этамбутол (EMB), изониазид (INH), пиразинамид (PZA), рифампицин (RIF), стрептомицин (SM), и 15 препаратов второй линии, таких как амикацин (AMK), амоксициллин – клавуланат (AMX – CL), капреомицин (CM), циклосерин (CS), этионамид (ETO), имипенем – циластатин (IMI), канамицин (KM), левофлоксацин (LFX), линезолид (LZD), моксифлоксацин (MFX), моксифлоксацин 0,25 (MFX 0,25), офлоксацин (OFX), параминосалициловая кислота (PAS), протионамид (PTH) и микобутин (рифабутин) (RFB). Эти данные

были взяты с портала по туберкулезу [24], представляющего собой платформу для исследований лекарственно-устойчивого туберкулеза, которая содержит проверенную информацию об устойчивости или чувствительности образцов МБТ к рассматриваемым препаратам. В табл. 1 представлено распределение чувствительных и устойчивых образцов для каждого препарата из набора данных.

Таблица 1

Характеристика набора данных

Table 1

Characterisation of the dataset

Препарат	Количество чувствительных геномов	Количество устойчивых геномов	Препарат	Количество чувствительных геномов	Количество устойчивых геномов
EMB	839	617	IMI	384	180
INH	438	977	KM	861	767
PZA	447	498	LFX	777	1188
RIF	627	933	LZD	719	766
SM	539	875	MFX	1159	413
AMK	942	1302	MFX 0,25	254	137
AMX – CL	626	297	OFX	589	385
CM	943	1185	PAS	705	1264
CS	810	576	PTH	605	1127
ETO	155	158	RFB	197	169

Меры качества предсказания. Качество предсказания фенотипа на основе ОНП в рассматриваемой позиции генома можно оценить по матрице ОНП и вектору фенотипов с помощью нескольких показателей, таких как

$$\begin{aligned}
 \text{точность} &= \frac{TP}{TP + FP}, \\
 \text{полнота} &= \frac{TP}{TP + FN}, \\
 \text{специфичность} &= \frac{TN}{TN + FP}, \\
 \text{правильность} &= \frac{TP + TN}{TP + TN + FP + FN},
 \end{aligned} \tag{2}$$

где TP – число истинно положительных предсказаний; TN – число истинно отрицательных предсказаний; FP – число ложноположительных предсказаний; FN – число ложноотрицательных предсказаний.

Прогнозирование лекарственной устойчивости основано на наличии соответствующего ОНП. Если образец устойчив к препарату d и предсказан как устойчивый, то результат прогнозирования считается истинно положительным. Аналогично, если образец чувствителен к препарату d и предсказан как чувствительный, то прогноз является истинно отрицательным. Однако ни одно предсказание не идеально. Так, если образец устойчив к препарату d , а прогнозируется, что он чувствителен, то такое предсказание считается ложноотрицательным. Аналогично, если образец чувствителен к препарату d , но предсказан как устойчивый, то результат прогнозирования будет ложноположительным.

Из формулы (2) следует, что точность – это отношение общего числа правильно предсказанных положительных образцов к общему числу образцов, предсказанных как положительные. Полнота представляет собой отношение общего числа правильно предсказанных положительных образцов к общему числу положительных образцов. Специфичность определяется как отношение общего числа правильно предсказанных отрицательных образцов к общему числу отрицательных образцов. Правильность отражает долю правильно предсказанных образцов, как положительных, так и отрицательных, от общего количества образцов.

Алгоритмы взаимодействия ОНП. Если исходная последовательность ОНП $\mathbf{x} = (x_1, x_2, \dots, x_p)$ содержит тысячи элементов, то можно вычислить значения меры правильности $a(x_i, x_j)$ (2) для комбинирования

рованных ОНП, образованных всеми парами одиночных ОНП (x_i, x_j) . В дальнейшем полученная информация используется для поиска наиболее важных комбинаций ОНП.

Выберем q пар (x_i, x_j) , $q \geq 1$, с наибольшими значениями $a(x_i, x_j)$, которые содержат различные ОНП. Продолжим отобранные пары ОНП до комбинаций, состоящих из не более чем l ОНП, $l \geq 2$, с помощью жадного алгоритма следующим образом. Для всех отобранных пар (x_i, x_j) рассмотрим комбинации трех ОНП (x_i, x_j, x_r) для каждого x_r из начальной последовательности x , отличного от x_i и x_j . Выберем произвольную комбинацию трех ОНП (x_i, x_j, x_k) с максимальным значением $a(x_i, x_j, x_k)$. Аналогичным образом можно получить комбинации, состоящие из не более чем l ОНП, начиная с отобранной пары (x_i, x_j) . Наконец, выберем либо один ОНП, либо комбинацию ОНП, ассоциированные с устойчивостью к рассматриваемому препарату, с максимальным значением меры правильности. Эта процедура может быть формализована алгоритмами 1 и 2 для аддитивного и мультипликативного режимов взаимодействия ОНП соответственно.

Алгоритм 1 (аддитивное взаимодействие ОНП).

Дана последовательность ОНП $x = (x_1, x_2, \dots, x_p)$, отсортированных в невозрастающем порядке значений меры правильности $a(x_1) \geq a(x_2) \geq \dots \geq a(x_p)$.

Шаг 1. Вычислим значения $a(x_i, x_j)$ для каждой пары ОНП, используя аддитивный режим взаимодействия.

Шаг 2. Выберем q пар ОНП с максимальными значениями меры правильности.

Шаг 3. Продолжим каждую выбранную пару до комбинации, содержащей не более l ОНП, с помощью жадного алгоритма для аддитивных взаимодействий.

Шаг 4. Выберем либо отдельный ОНП, либо комбинацию не более чем l ОНП, связанные с устойчивостью к рассматриваемому препарату, с максимальным значением меры правильности.

Алгоритм 2 (мультипликативное взаимодействие ОНП).

Дана последовательность ОНП $x = (x_1, x_2, \dots, x_p)$, отсортированных в невозрастающем порядке значений меры правильности $a(x_1) \geq a(x_2) \geq \dots \geq a(x_p)$.

Шаг 1. Вычислим значения $a(x_i, x_j)$ для каждой пары ОНП, используя мультипликативный режим взаимодействия.

Шаг 2. Выберем q пар ОНП с максимальными значениями меры правильности.

Шаг 3. Продолжим каждую выбранную пару до комбинации, содержащей не более l ОНП, с помощью жадного алгоритма для мультипликативных взаимодействий.

Шаг 4. Выберем либо отдельный ОНП, либо комбинацию не более чем l ОНП, связанные с устойчивостью к рассматриваемому препарату, с максимальным значением меры правильности.

Выбор режима взаимодействия ОНП влияет на прогностические свойства комбинированных мутаций. Аддитивные взаимодействия имеют тенденцию к уменьшению числа истинно отрицательных и увеличению числа ложноположительных предсказаний, в то время как мультипликативные взаимодействия имеют тенденцию к увеличению числа ложноотрицательных и уменьшению числа истинно положительных предсказаний. В связи с этим представляется разумным использовать смешанный режим аддитивного и мультипликативного взаимодействий. Рассмотрим два дополнительных алгоритма. Алгоритм 3 применяет мультипликативное взаимодействие для продолжения пар ОНП, полученных с помощью аддитивного взаимодействия, а алгоритм 4 использует аддитивное взаимодействие для продолжения пар ОНП, полученных с помощью мультипликативного взаимодействия.

Алгоритм 3 (аддитивно-мультипликативное взаимодействие ОНП).

Дана последовательность ОНП $x = (x_1, x_2, \dots, x_p)$, отсортированных в невозрастающем порядке значений меры правильности $a(x_1) \geq a(x_2) \geq \dots \geq a(x_p)$.

Шаг 1. Вычислим значения $a(x_i, x_j)$ для каждой пары ОНП, используя аддитивный режим взаимодействия.

Шаг 2. Выберем q пар ОНП с максимальными значениями меры правильности.

Шаг 3. Продолжим каждую выбранную пару до комбинации, содержащей не более l ОНП, с помощью жадного алгоритма с использованием мультипликативных взаимодействий.

Шаг 4. Выберем либо отдельный ОНП, либо комбинацию не более чем l ОНП, связанные с устойчивостью к рассматриваемому препарату, с максимальным значением меры правильности.

Алгоритм 4 (мультипликативно-аддитивное взаимодействие ОНП).

Дана последовательность ОНП $\mathbf{x} = (x_1, x_2, \dots, x_p)$, отсортированных в невозрастающем порядке значений меры правильности $a(x_1) \geq a(x_2) \geq \dots \geq a(x_p)$.

Шаг 1. Вычислим значения $a(x_i, x_j)$ для каждой пары ОНП, используя мультипликативный режим взаимодействия.

Шаг 2. Выберем q пар ОНП с максимальными значениями меры правильности.

Шаг 3. Продолжим каждую выбранную пару до комбинации, содержащей не более l ОНП, с помощью жадного алгоритма с использованием аддитивных взаимодействий.

Шаг 4. Выберем либо отдельный ОНП, либо комбинацию не более чем l ОНП, связанные с устойчивостью к рассматриваемому препарату, с максимальным значением меры правильности.

Если длина последовательности $\mathbf{x}$ измеряется сотнями тысяч элементов, то прямая проверка значений меры правильности для всех пар (x_i, x_j) становится невозможной из-за ограниченности вычислительных ресурсов. В этом случае можно выбрать подпоследовательность исходной последовательности $\mathbf{x}$, содержащую разумное количество мутаций, которое еще позволяет вычислить значения используемой меры качества предсказания для всех пар ее элементов, а затем, следуя описанным выше процедурам, получить комбинации не более чем l ОНП с помощью жадного алгоритма. Понятно, что все алгоритмы могут быть использованы и для других мер качества предсказания (2).

Результаты и их обсуждение

Применение алгоритмов к препаратам первой линии. Для снижения вычислительной сложности для каждого препарата первой линии отобраны 25 пар ОНП ($q = 25$) с наибольшим значением меры правильности, которые были продолжены с помощью разработанных алгоритмов до комбинированных ОНП, содержащих не более 5 мутаций ($l = 5$). Для препаратов EMB, INH, PZA и SM максимальные значения меры правильности (79,5; 90,8; 75,8 и 82,1 % соответственно) получены с помощью алгоритма 1 (табл. 2). Для препарата RIF наибольшая величина меры правильности (88,9 %) была достигнута с использованием алгоритма 4. Из рис. 1 видно, что комбинации двух ОНП, использующие либо аддитивный, либо мультипликативный режим взаимодействия, дают лучшие значения рассматриваемой меры качества предсказания, чем одиночные ОНП.

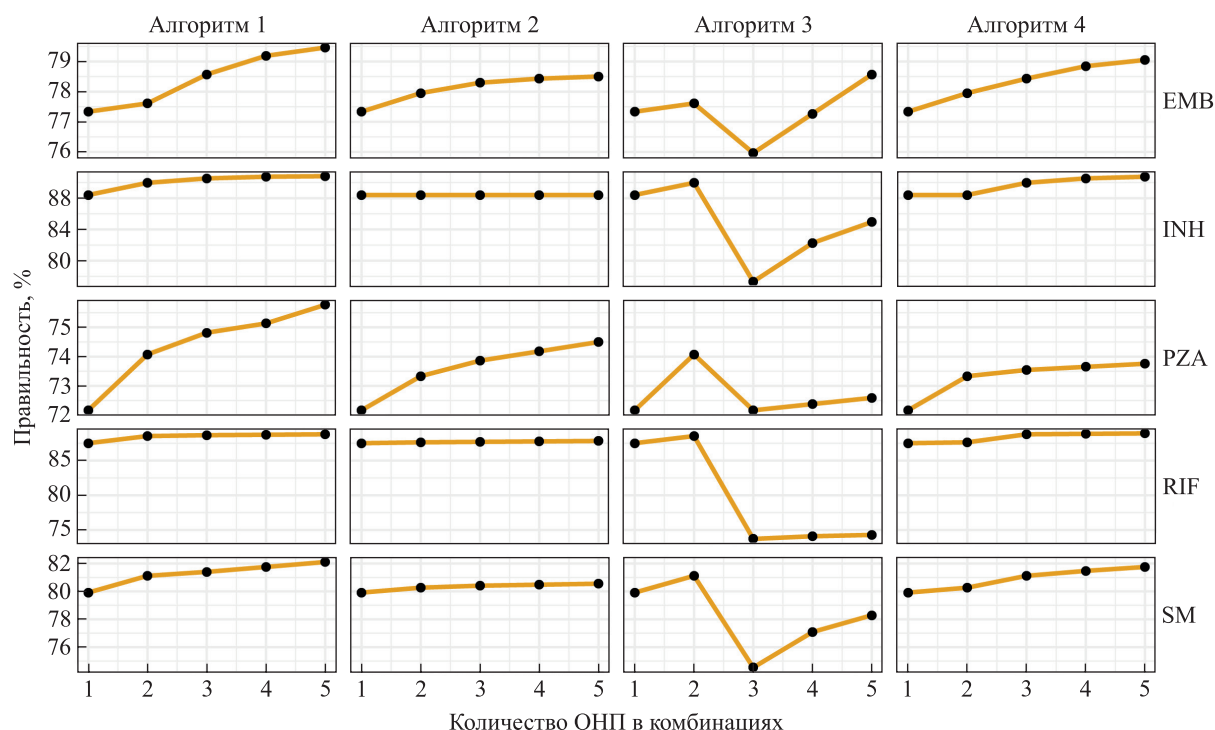


Рис. 1. Комбинации l ОНП, $l \in \{2, 3, 4, 5\}$, наиболее значимо ассоциированные с устойчивостью к препаратам первой линии
Fig. 1. Combinations of l single nucleotide polymorphisms (SNPs), $l \in \{2, 3, 4, 5\}$, most significantly associated with resistance to the first-line drugs

Таблица 2

Правильность прогнозирования устойчивости
к препаратам первой линии

Table 2

Accuracy of predicting resistance
to the first-line drugs

Препарат	Количество предсказаний				Правильность, %
	TN	FP	FN	TP	
Алгоритм 1					
EMB	590	249	50	567	79,5
INH	401	37	93	884	90,8
PZA	292	155	74	424	75,8
RIF	504	123	52	881	88,8
SM	371	168	85	790	82,1
Алгоритм 2					
EMB	587	252	61	556	78,5
INH	423	15	149	828	88,4
PZA	259	188	53	445	74,5
RIF	523	104	86	847	87,8
SM	382	157	118	757	80,6
Алгоритм 3					
EMB	597	242	70	547	78,6
INH	421	17	125	852	90,0
PZA	269	178	67	431	74,1
RIF	508	119	60	873	88,5
SM	354	185	82	793	81,1
Алгоритм 4					
EMB	573	266	39	578	79,1
INH	402	36	95	882	90,7
PZA	223	224	24	474	73,8
RIF	514	113	60	873	88,9
SM	358	181	77	798	81,8

Примечание. Полу жирным начертанием выделены наилучшие значения меры правильности для каждого препарата.

Применение алгоритмов к препаратам второй линии. Значения меры правильности предсказания, полученные с помощью рассматриваемых алгоритмов для препаратов второй линии, представлены в табл. 3. В большинстве случаев лучшие результаты дает алгоритм 4, для 5 препаратов – алгоритм 1, для 3 препаратов – алгоритм 2. Почти везде наибольшая величина меры правильности предсказания по двум ОНП превышает лучший результат, достигнутый при использовании одного ОНП (рис. 2). По мере увеличения количества ОНП в комбинациях ($l \in \{2, 3, 4, 5\}$) максимальное значение меры правильности предсказания изменяется (см. рис. 2).

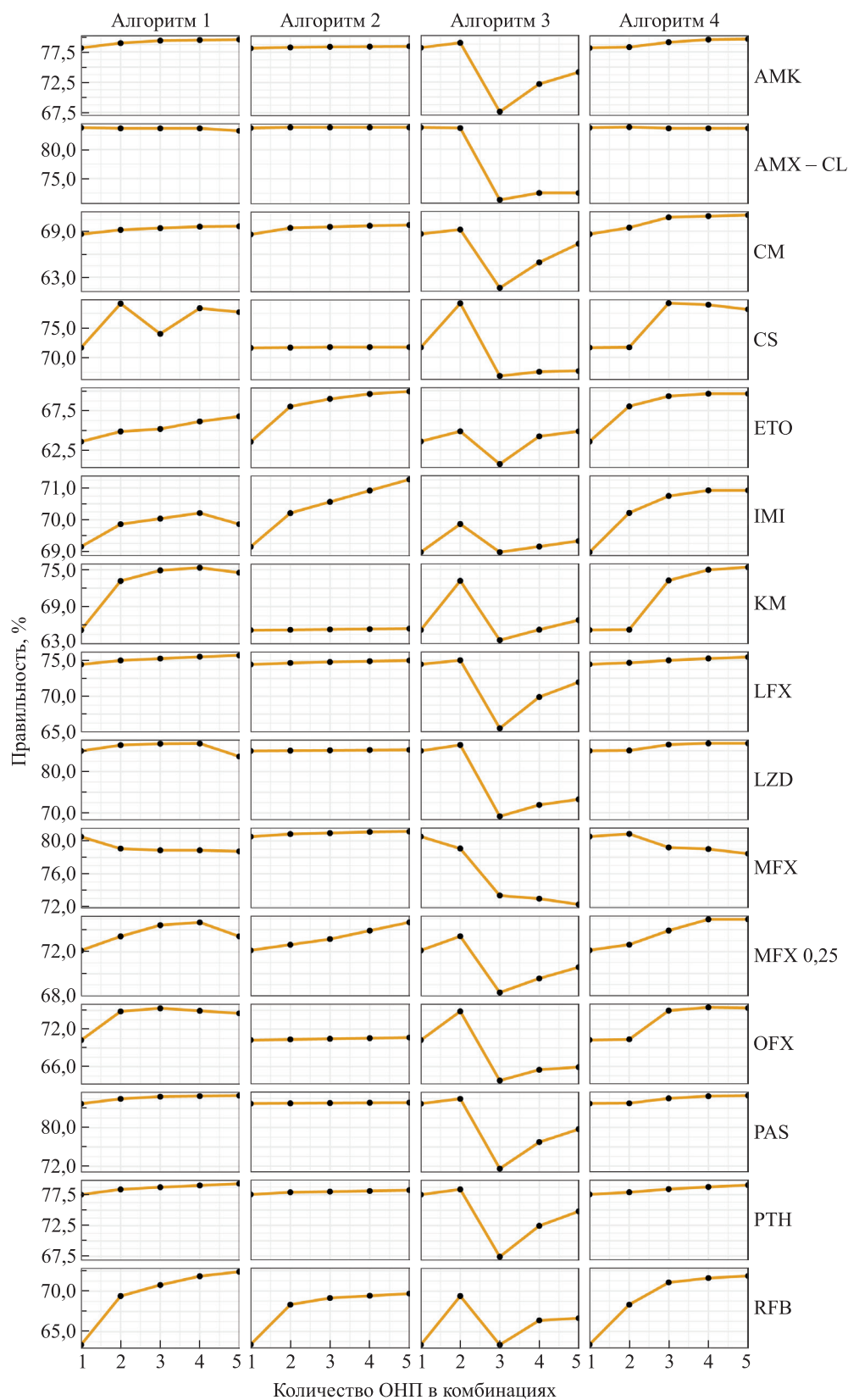


Рис. 2. Комбинации l ОНП, $l \in \{2, 3, 4, 5\}$,
наиболее значимо ассоциированные с устойчивостью к препаратам второй линии

Fig. 2. Combinations of l SNPs, $l \in \{2, 3, 4, 5\}$,
most significantly associated with resistance to the second-line drugs

Таблица 3

Правильность прогнозирования устойчивости
к препаратам второй линии

Table 3

Accuracy of predicting resistance
to the second-line drugs

Препарат	Количество предсказаний				Правильность, %
	TN	FP	FN	TP	
Алгоритм 1					
AMK	549	393	66	1236	79,5
AMX – CL	592	34	116	181	83,7
CM	384	559	87	1098	69,6
CS	739	71	218	358	79,1
ETO	83	72	34	124	66,1
IMI	322	62	106	74	70,2
KM	689	172	230	537	75,3
LFX	382	395	82	1106	75,7
LZD	575	144	52	714	86,8
MFX	1062	97	209	204	80,5
MFX 0,25	204	50	49	88	74,7
OFX	525	64	176	209	75,4
PAS	505	200	63	1201	86,6
PTH	332	273	85	1042	79,3
RFB	133	64	37	132	72,4
Алгоритм 2					
AMK	567	375	108	1194	78,5
AMX – CL	593	33	116	181	83,9
CM	943	0	1185	0	44,3
CS	810	0	576	0	58,4
ETO	155	0	158	0	49,5
IMI	376	8	154	26	71,3
KM	813	48	515	252	65,4
LFX	428	349	142	1046	75,0
LZD	583	136	83	683	85,3
MFX	1078	81	215	198	81,2
MFX 0,25	235	19	80	57	74,7
OFX	568	21	265	120	70,6
PAS	705	0	1264	0	35,8
PTH	361	244	133	994	78,2
RFB	155	42	69	100	69,7
Алгоритм 3					
AMK	549	393	79	1223	79,0
AMX – CL	592	34	116	181	83,7
CM	404	539	117	1068	69,2
CS	739	71	218	358	79,1
ETO	65	90	20	138	64,9

Окончание табл. 3
Ending of the table 3

Препарат	Количество предсказаний				Правильность, %
	TN	FP	FN	TP	
IMI	345	39	133	47	69,5
KM	773	88	349	418	73,2
LFX	388	389	102	1086	75,0
LZD	576	143	59	707	86,4
MFX	1062	97	209	204	80,5
MFX 0,25	228	26	78	59	73,4
OFX	549	40	205	180	74,8
PAS	505	200	76	1188	86,0
PTH	335	270	104	1023	78,4
RFB	126	71	41	128	69,4
<i>Алгоритм 4</i>					
AMK	553	389	68	1234	79,6
AMX – CL	593	33	116	181	83,9
CM	500	443	172	1013	71,1
CS	740	70	218	358	79,2
ETO	85	70	25	133	69,6
IMI	342	42	122	58	70,9
KM	690	171	230	537	75,4
LFX	382	395	87	1101	75,5
LZD	577	142	54	712	86,8
MFX	1072	87	214	199	80,9
MFX 0,25	224	30	68	69	74,9
OFX	526	63	176	209	75,5
PAS	506	199	65	1199	86,6
PTH	332	273	90	1037	79,0
RFB	116	81	22	147	71,9

Примечание. Полужирным начертанием выделены наилучшие значения меры правильности для каждого препарата.

Сравнение с другими алгоритмами. Эффективность разработанных алгоритмов была сопоставлена с эффективностью программных систем *Mykrobe* [15] и *TB-Profiler* [20], которые относятся к числу лучших вычислительных инструментов для предсказания лекарственной устойчивости МБТ. В системе *Mykrobe* используется граф де Брейна для представления генома и байесовский вывод для предсказания устойчивости к лекарственным препаратам. Система *TB-Profiler* строит прогноз устойчивости к лекарственным препаратам на основе анализа ОНП и набора правил для их интерпретации.

Значения меры правильности, полученные с помощью программ *Mykrobe* (версия 0.13.0) и *TB-Profiler* (версия 5.0.1) для 13 препаратов, были вычислены с использованием того же набора данных, который применялся в разработанных авторами алгоритмах. Результаты сравнения алгоритмов 1–4 с программными системами *Mykrobe* и *TB-Profiler* приведены в табл. 4, анализ которой свидетельствует о том, что для всех препаратов первой линии и препарата OFX второй линии системы *Mykrobe* и *TB-Profiler* по правильности прогнозирования незначительно превосходят алгоритмы 1–4, однако для остальных препаратов второй линии уступают им. Дополнительным преимуществом предложенных авторами алгоритмов является интуитивно понятная интерпретация полученных результатов, поскольку устойчивость к конкретному препарату предсказывается по наличию соответствующих индивидуальных мутаций в геноме образца и по режиму их взаимодействия.

Таблица 4

Сравнение разработанных алгоритмов с системами *Mykrobe* и *TB-Profiler*
для прогнозирования устойчивости к противотуберкулезным препаратам первой и второй линий

Table 4

Comparison of the developed algorithms with *Mykrobe* and *TB-Profiler* systems
for prediction of resistance to the first-line and second-line anti-tuberculosis drugs

Препарат	Количество предсказаний				Полнота, %	Специ- фичность, %	Точность, %	Правиль- ность, %	Алгоритм или программа
	TN	FP	FN	TP					
EMB	590	249	50	567	91,9	70,3	69,5	79,5	Алгоритм 1
	661	178	83	534	86,5	78,8	75,0	82,1	<i>Mykrobe</i>
	604	235	42	575	93,2	72,0	71,0	81,0	<i>TB-Profiler</i>
INH	401	37	93	884	90,5	91,6	96,0	90,8	Алгоритм 1
	419	19	104	873	89,4	95,7	97,9	91,3	<i>Mykrobe</i>
	417	21	99	878	89,9	95,2	97,7	91,5	<i>TB-Profiler</i>
PZA	292	155	74	424	85,1	65,3	73,2	75,8	Алгоритм 1
	374	73	101	397	79,7	83,7	84,5	81,6	<i>Mykrobe</i>
	361	86	87	411	82,5	80,8	82,7	81,7	<i>TB-Profiler</i>
RIF	514	113	60	873	93,6	82,0	88,5	88,9	Алгоритм 4
	532	95	69	864	92,6	84,8	90,1	89,5	<i>Mykrobe</i>
	530	97	58	875	93,8	84,5	90,0	90,1	<i>TB-Profiler</i>
SM	371	168	85	790	90,3	68,8	82,5	82,1	Алгоритм 1
	349	190	46	829	94,7	64,7	81,4	83,3	<i>Mykrobe</i>
	357	182	61	814	93,0	66,2	81,7	82,8	<i>TB-Profiler</i>
AMK	553	389	68	1234	94,8	58,7	76,0	79,6	Алгоритм 3
	898	44	878	424	32,6	95,3	90,6	58,9	<i>Mykrobe</i>
	895	47	870	432	33,2	95,0	90,2	59,1	<i>TB-Profiler</i>
AMX – CL	593	33	116	181	60,9	94,7	84,6	83,9	Алгоритм 2
	Неприменимо								<i>Mykrobe</i>
	Неприменимо								<i>TB-Profiler</i>
CM	500	443	172	1013	85,5	53,0	69,6	71,1	Алгоритм 4
	899	44	838	347	29,3	95,3	88,7	58,6	<i>Mykrobe</i>
	894	49	794	391	33,0	94,8	88,9	60,4	<i>TB-Profiler</i>
CS	740	70	218	358	62,2	91,4	83,6	79,2	Алгоритм 4
	Неприменимо								<i>Mykrobe</i>
	788	22	518	58	10,1	97,3	72,5	61,0	<i>TB-Profiler</i>
ETO	85	70	25	133	84,2	54,8	65,5	69,6	Алгоритм 4
	76	79	44	114	72,2	49,0	59,1	60,7	<i>Mykrobe</i>
	70	85	36	122	77,2	45,2	58,9	61,3	<i>TB-Profiler</i>
IMI	376	8	154	26	14,4	97,9	76,5	71,3	Алгоритм 2
	Неприменимо								<i>Mykrobe</i>
	Неприменимо								<i>TB-Profiler</i>
KM	690	171	230	537	70,0	80,1	75,8	75,4	Алгоритм 4
	586	275	168	599	78,1	68,1	68,5	72,8	<i>Mykrobe</i>
	573	288	155	612	79,8	66,6	68,0	72,8	<i>TB-Profiler</i>

Окончание табл. 4
Ending of the table 4

Препарат	Количество предсказаний				Полнота, %	Специ- фичность, %	Точность, %	Правиль- ность, %	Алгоритм или программа
	TN	FP	FN	TP					
LFX	382	395	82	1106	93,1	49,2	73,7	75,7	Алгоритм 1
	699	78	588	600	50,5	90,0	88,5	66,1	<i>Mykrobe</i>
	695	82	552	636	53,5	89,4	88,6	67,7	<i>TB-Profiler</i>
LZD	575	144	52	714	93,2	80,0	83,2	86,8	Алгоритм 1
	719	0	764	2	0,3	100,0	100,0	48,6	<i>Mykrobe</i>
	719	0	763	3	0,4	100,0	100,0	48,6	<i>TB-Profiler</i>
MFX	1078	81	215	198	47,9	93,0	71,0	81,2	Алгоритм 2
	783	376	90	323	78,2	67,6	46,2	70,4	<i>Mykrobe</i>
	753	406	80	333	80,6	65,0	45,1	69,1	<i>TB-Profiler</i>
MFX 0,25	224	30	68	69	50,4	88,2	69,7	74,9	Алгоритм 4
	Неприменимо								<i>Mykrobe</i>
	Неприменимо								<i>TB-Profiler</i>
OFX	526	63	176	209	54,3	89,3	76,8	75,5	Алгоритм 4
	535	54	110	275	71,4	90,8	83,6	83,2	<i>Mykrobe</i>
	525	64	99	286	74,3	89,1	81,7	83,3	<i>TB-Profiler</i>
PAS	505	200	63	1201	95,0	71,6	85,7	86,6	Алгоритм 1
	Неприменимо								<i>Mykrobe</i>
	690	15	1149	115	9,1	97,9	88,5	40,9	<i>TB-Profiler</i>
PTH	332	273	85	1042	92,5	54,9	79,2	79,3	Алгоритм 1
	Неприменимо								<i>Mykrobe</i>
	Неприменимо								<i>TB-Profiler</i>
RFB	133	64	37	132	78,1	67,5	67,3	72,4	Алгоритм 1
	Неприменимо								<i>Mykrobe</i>
	Неприменимо								<i>TB-Profiler</i>

Примечание. Полу жирным начертанием выделены наилучшие значения меры правильности для каждого препарата и алгоритм либо программа, с помощью которых они были получены.

В табл. 5 приведены ОНП, показавшие наилучшие результаты, и соответствующие им гены. В большинстве случаев эти ОНП распределены по разным генам, и только у некоторых препаратов несколько важных ОНП находятся в пределах одного гена.

Таблица 5

Важные ОНП и соответствующие им гены

Table 5

Important SNPs and corresponding genes

Препарат	ОНП	Гены	Правильность, %
<i>Первая линия</i>			
EMB	rs4247429	<i>embB</i>	79,5
	rs761155	<i>rpoB</i>	
	rs4248003	<i>embB</i>	
	rs4247431	<i>embB</i>	
	rs2101193	<i>ureD</i>	

Продолжение табл. 5
Continuation of the table 5

Препарат	ОНП	Гены	Правильность, %
INH	rs2155168	<i>katG</i>	90,8
	rs1673425	NA	
	rs761155	<i>rpoB</i>	
	rs781687	<i>rpsL</i>	
	rs184866	<i>pntAb</i>	
PZA	rs2904110	<i>dhaA</i>	75,8
	rs781822	<i>rpsL</i>	
	rs4247429	<i>embB</i>	
	rs4248003	<i>embB</i>	
	rs4247431	<i>embB</i>	
RIF	rs1977	NA	88,9
	rs2155168	<i>katG</i>	
	rs761155	<i>rpoB</i>	
	rs7582	<i>gyrA</i>	
	rs1473246	<i>rrs</i>	
SM	rs1472359	<i>rrs</i>	82,1
	rs761155	<i>rpoB</i>	
	rs781687	<i>rpsL</i>	
	rs781822	<i>rpsL</i>	
	rs184866	<i>pntAb</i>	
Вторая линия			
AMK	rs1093406	<i>PE_PGRS17</i>	79,6
	rs2155168	<i>katG</i>	
	rs1673425	NA	
	rs1473246	<i>rrs</i>	
	rs4247431	<i>embB</i>	
AMX – CL	rs2923391	<i>ruvB</i>	83,9
	rs1473246	<i>rrs</i>	
CM	rs2155168	<i>katG</i>	71,1
	rs3590686	NA	
	rs1473246	<i>rrs</i>	
	rs337976	<i>PE_PGRS4</i>	
	rs4247431	<i>embB</i>	
CS	rs7582	<i>gyrA</i>	79,2
	rs1933988	NA	
	rs7570	<i>gyrA</i>	
ETO	rs840496	<i>PE_PGRS10</i>	69,6
	rs761155	<i>rpoB</i>	
	rs104962	<i>Rv0095c</i>	
	rs2045310	<i>PE_PGRS32</i>	
	rs24698	<i>fhaA</i>	

Продолжение табл. 5
Continuation of the table 5

Препарат	ОНП	Гены	Правильность, %
IMI	rs2944769	<i>PE_PGRS45</i>	71,3
	rs2196855	<i>Rv1945</i>	
	rs336005	<i>Rv0278c</i>	
	rs71336	<i>Rv0064</i>	
	rs3732624	<i>PPE54</i>	
KM	rs7582	<i>gyrA</i>	75,4
	rs1093406	<i>PE_PGRS17</i>	
	rs7570	<i>gyrA</i>	
	rs1473246	<i>rrs</i>	
	rs493324	<i>pta</i>	
LFX	rs2155168	<i>katG</i>	75,7
	rs781687	<i>rpsL</i>	
	rs781822	<i>rpsL</i>	
	rs1472359	<i>rrs</i>	
	rs1473246	<i>rrs</i>	
LZD	rs2155168	<i>katG</i>	86,8
	rs761155	<i>rpoB</i>	
	rs1473246	<i>rrs</i>	
	rs761139	<i>rpoB</i>	
MFX	rs1473246	<i>rrs</i>	81,2
	rs1224367	NA	
	rs1499274	<i>pncB1</i>	
	rs3746409	<i>PPE55</i>	
	rs177857	<i>PE1</i>	
MFX 0,25	rs7582	<i>gyrA</i>	74,9
	rs3883796	<i>Rv3466</i>	
	rs3738503	<i>PE_PGRS50</i>	
	rs2715344	NA	
OFX	rs7582	<i>gyrA</i>	75,5
	rs92199	<i>Rv0083</i>	
	rs7570	<i>gyrA</i>	
	rs2030634	<i>esxM</i>	
PAS	rs2155168	<i>katG</i>	86,6
	rs1673425	NA	
	rs1473246	<i>rrs</i>	
	rs4247429	<i>embB</i>	
	rs4247431	<i>embB</i>	

Окончание табл. 5
Ending of the table 5

Препарат	ОНП	Гены	Правильность, %
PTH	rs2155168	<i>katG</i>	79,3
	rs781687	<i>rpsL</i>	
	rs781822	<i>rpsL</i>	
	rs1472359	<i>rrs</i>	
	rs1473246	<i>rrs</i>	
RFB	rs1626935	<i>zwf2</i>	72,4
	rs1130195	<i>ispE</i>	
	rs4243221	NA	
	rs4247429	<i>embB</i>	
	rs1161026	<i>esxJ</i>	

Примечание. Аббревиатура NA означает, что данный ОНП не находится в гене.

Заключение

Таким образом, в данном исследовании предложены четыре новых алгоритма для выявления комбинаций индивидуальных мутаций, ассоциированных с устойчивостью патогенных организмов к лекарственным препаратам. Разработанные алгоритмы основаны на анализе геномов и концепции взаимодействия мутаций, предполагающей, что комбинации нескольких ОНП позволяют предсказывать фенотип с более высоким значением меры правильности, чем одиночные ОНП. Рассмотрены два режима взаимодействия мутаций – аддитивный и мультипликативный. Предложенные алгоритмы используют различные последовательности режимов взаимодействия ОНП. Важным моментом является то, что все алгоритмы на первом этапе применяют полный перебор пар мутаций для поиска тех из них, которые демонстрируют наибольшие значения меры правильности предсказания устойчивости к лекарству. На втором этапе отобранные пары ОНП используются для получения комбинаций, содержащих до 5 мутаций, с помощью жадного алгоритма. Очевидно, что исчерпывающий поиск пар ОНП возможен лишь при ограничении числа отдельных ОНП, делающем процедуру вычислительно выполнимой. В случае прогнозирования устойчивости МБТ к лекарственным препаратам максимальное количество ОНП, которые необходимо рассмотреть, составляет менее 10 000, поэтому исчерпывающий поиск позволил проверить все их пары. В итоге удалось выявить комбинации, содержащие до 5 мутаций и ассоциированные с устойчивостью к 20 препаратам, применяемым в настоящее время для лечения туберкулеза. В будущем авторы планируют усовершенствовать предложенный подход и найти комбинации ОНП с более высокими величинами ассоциации с лекарственной устойчивостью МБТ.

Библиографические ссылки

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*. 2017;101(1):5–22. DOI: 10.1016/j.ajhg.2017.06.005.
2. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: realizing the promise. *The American Journal of Human Genetics*. 2023;110(2):179–194. DOI: 10.1016/j.ajhg.2022.12.011.
3. Zhu Zhixiang, Tong Xiaoran, Zhu Zhihong, Liang Meimei, Cui Wenyan, Su Kunkai, et al. Development of GMDR-GPU for gene – gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes. *PloS One*. 2013;8(4):e61943. DOI: 10.1371/journal.pone.0061943.
4. Díez Díaz F, Sánchez Lasheras F, Moreno V, Moratalla-Navarro F, Molina de la Torre AJ, Martín Sánchez V. GASVeM: a new machine learning methodology for multi-SNP analysis of GWAS data based on genetic algorithms and support vector machines. *Mathematics*. 2021;9(6):654. DOI: 10.3390/math9060654.
5. Nguyen T-T, Huang JZ, Wu Q, Nguyen TT, Li MJ. Genome-wide association data classification and SNPs selection using two-stage quality-based random forests. *BMC Genomics*. 2015;16(supplement 2):S5. DOI: 10.1186/1471-2164-16-S2-S5.
6. Ritchie MD, Van Steen K. The search for gene – gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. *Annals of Translational Medicine*. 2018;6(8):157. DOI: 10.21037/atm.2018.04.05.

7. Berrandou T-E, Balding D, Speed D. LDAK-GBAT: fast and powerful gene-based association testing using summary statistics. *The American Journal of Human Genetics*. 2023;110(1):23–29. DOI: 10.1016/j.ajhg.2022.11.010.
8. Zhang J, Liang X, Gonzales S, Liu J, Gao XR, Wang X. A gene based combination test using GWAS summary data. *BMC Bioinformatics*. 2023;24:2. DOI: 10.1186/s12859-022-05114-x.
9. Evans LM, Arehart CH, Grotzinger AD, Mize TJ, Brasher MS, Stitzel JA, et al. Transcriptome-wide gene – gene interaction associations elucidate pathways and functional enrichment of complex traits. *PLoS Genetics*. 2023;19(5):e1010693. DOI: 10.1371/journal.pgen.1010693.
10. World Health Organization. *Global tuberculosis report – 2023*. Geneva: World Health Organization; 2023. XIV, 57 p.
11. Goossens SN, Sampson SL, Van Rie A. Mechanisms of drug-induced tolerance in *Mycobacterium tuberculosis*. *Clinical Microbiology Reviews*. 2020;34(1):e00141-20. DOI: 10.1128/cmr.00141-20.
12. Islam MM, Hameed HMA, Mugweru J, Chhotaray C, Wang C, Tan Y, et al. Drug resistance mechanisms and novel drug targets for tuberculosis therapy. *Journal of Genetics and Genomics*. 2017;44(1):21–37. DOI: 10.1016/j.jgg.2016.10.002.
13. Cui Z-J, Yang Q-Y, Zhang H-Y, Zhu Q, Zhang Q-Y. Bioinformatics identification of drug resistance-associated gene pairs in *Mycobacterium tuberculosis*. *International Journal of Molecular Sciences*. 2016;17(9):1417. DOI: 10.3390/ijms17091417.
14. Kuang X, Wang F, Hernandez KM, Zhang Z, Grossman RL. Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN. *Scientific Reports*. 2022;12:2427. DOI: 10.1038/s41598-022-06449-4.
15. Hunt M, Bradley P, Lapierre SG, Heys S, Thomsit M, Hall MB, et al. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Research*. 2019;4:191. DOI: 10.12688/wellcomeopenres.15603.1.
16. Benavente ED, Coll F, Furnham N, McNerney R, Glynn JR, Campino S, et al. PhyTB: phylogenetic tree visualisation and sample positioning for *M. tuberculosis*. *BMC Bioinformatics*. 2015;16:155. DOI: 10.1186/s12859-015-0603-3.
17. Ezewudo M, Borens A, Chiner-Oms Á, Miotto P, Chindelevitch L, Starks AM, et al. Integrating standardized whole genome sequence analysis with a global *Mycobacterium tuberculosis* antibiotic resistance knowledgebase. *Scientific Reports*. 2018;8:15382. DOI: 10.1038/s41598-018-33731-1.
18. Gröschel M, Owens M, Freschi L, Vargas R Jr, Marin MG, Phelan J, et al. GenTB: a user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Medicine*. 2021;13:138. DOI: 10.1186/s13073-021-00953-4.
19. Iwai H, Kato-Miyazawa M, Kirikae T, Miyoshi-Akiyama T. CASTB (the comprehensive analysis server for the *Mycobacterium tuberculosis* complex): a publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. *Tuberculosis*. 2015;95(6):843–844. DOI: 10.1016/j.tube.2015.09.002.
20. Phelan JE, O’Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Medicine*. 2019;11:41. DOI: 10.1186/s13073-019-0650-x.
21. Sekizuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S, et al. TGS-TB: total genotyping solution for *Mycobacterium tuberculosis* using short-read whole-genome sequencing. *PloS One*. 2015;10(11):e0142951. DOI: 10.1371/journal.pone.0142951.
22. Sergeev RS, Kavaliou IS, Sataneuski UV, Gabrielian A, Rosenthal A, Tartakovsky M, et al. Genome-wide analysis of MDR and XDR tuberculosis from Belarus: machine-learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2019;16(4):1398–1408. DOI: 10.1109/TCBB.2017.2720669.
23. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. *KvarQ*: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*. 2014;15:881. DOI: 10.1186/1471-2164-15-881.
24. Rosenthal A, Gabrielian A, Engle E, Hurt DE, Alexandru S, Crudu V, et al. The TB Portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis. *Journal of Clinical Microbiology*. 2017;55(11):3267–3282. DOI: <https://doi.org/10.1128/jcm.01013-17>.

Получена 25.08.2024 / исправлена 16.10.2024 / принята 16.10.2024.
Received 25.08.2024 / revised 16.10.2024 / accepted 16.10.2024.