УДК 004.021

Р.С. Сергеев[1], И.С. Ковалев[2], А.В. Тузиков[1], А. Розенталь[3], А. Габриэлян[3]

АЛГОРИТМЫ ПОИСКА МУТАЦИЙ ЛЕКАРСТВЕННОЙ УСТОЙЧИВОСТИ В ГЕНОМАХ МИКОБАКТЕРИЙ ТУБЕРКУЛЕЗА

Предлагается методология полногеномного поиска ассоциаций на примере геномов возбудителей туберкулеза, а также исследуются различные алгоритмы и подходы к оценке вклада мутаций в возникновение и развитие лекарственной устойчивости. Приводятся результаты экспериментов по поиску мутаций лекарственной устойчивости к основным противотуберкулезным препаратам на основании данных о пациентах из Беларуси.

Введение

Стремительное развитие высокопроизводительных методов секвенирования для определения нуклеотидных последовательностей ДНК живых организмов придает значительный импульс биологическим исследованиям и становлению персонализированной медицины. Однако сам по себе прочитанный генетический код не имеет большой практической ценности до извлечения из него нужной информации. Анализ расшифрованных геномных последовательностей зачастую приводит к задачам большой размерности, где число неизвестных параметров измеряется десятками тысяч при относительно небольшом числе доступных наблюдений. Одной из таких задач является поиск мутаций в геномах микроорганизмов, которые непосредственно обусловливают развитие лекарственной устойчивости к применяемым в ходе терапии лекарственным препаратам.

Актуальность задачи состоит в том, что в последнее время все чаще выявляются штаммы Mycobacterium tuberculosis, обладающие множественной и широкой лекарственной устойчивостью. Причина возникновения подобных штаммов во многом обусловлена такими факторами, как выбор неверной схемы лечения, нарушение режима химиотерапии, поздняя постановка диагноза [1]. Эти ошибки приводят к последовательному накоплению изменений в генах, прямо или косвенно вовлеченных во взаимодействие с лекарственными субстанциями.

Предлагаемая методология позволяет по набору полных геномов M. tuberculosis проанализировать влияние мутаций на развитие резистентности к противомикробным препаратам и на основании этой информации выполнить отбор участков генома для формирования базы данных структурных полиморфизмов. Исследование генетических или геномных маркеров лекарственной устойчивости имеет большое значение для своевременной диагностики форм туберкулеза, а также понимания биологических механизмов становления лекарственной устойчивости и выбора потенциальных мишеней для создания новых противомикробных препаратов.

1. Постановка задачи

В настоящем исследовании решается задача полногеномного поиска ассоциаций (genome-wide association study, GWAS), где анализируются мутации – однонуклеотидные полиморфизмы (single nucleotide polymorphisms, SNPs) в последовательностях ДНК микобактерий туберкулеза. Цель этой задачи состоит в идентификации участков генома, мутации в которых максимально влияют на фенотип (наличие либо отсутствие лекарственной устойчивости к определенному препарату).

Пусть на входе имеется набор геномных последовательностей S_1, S_2,...,S_n, таких, что $S_i = s_{i1}s_{i2}...s_{il}$, $s_{ij} \in \{A,T,G,C,-\}$. Здесь *A, T, G, C* обозначают четыре нуклеотида, а «–» означает, что элемент последовательности не определен. Будем считать, что все последовательности выравнены относительно друг друга и некоторой референсной последовательности S_0. Тогда лю-

бое отличие символа s_{ij} в i-й последовательности от символа s_{oj} в этой же позиции референсной последовательности будем называть однонуклеотидным полиморфизмом (точечной мутацией). Для более компактного представления введем матрицу генотипов X размера $n \times m$, строки которой будут соответствовать последовательностям, а в столбцы попадут только позиции исходного выравнивания, содержащие хотя бы одну мутацию. При этом $x_{ik} = 1$, если в i-й последовательности присутствует мутация в позиции, соответствующей k-му столбцу матрицы X, и $x_{ik} = 0$, если мутация отсутствует.

Существует несколько линий лекарственных препаратов, использующихся для терапии туберкулеза. Препараты первой линии назначаются в комбинациях для лечения первично инфицированных пациентов, у которых не обнаружено резистентных штаммов M. tuberculosis. Препараты второй линии применяются в случае, если установлена лекарственная устойчивость к препаратам первой линии. Информацию о результатах тестов на чувствительность к некоторому лекарственному препарату будем кодировать в виде вектора Y длины n, элемент которого $y_i = 1$, если установлена лекарственная устойчивость i-го организма к этому препарату, и $y_i = 0$ в противном случае.

В работе сравниваются ДНК-последовательности, выделенные из двух групп организмов: лекарственно-устойчивых и лекарственно-чувствительных. В качестве референсной последовательности S_0 был выбран геном штамма H37Rv (который сохраняет полную вирулентность и восприимчив к противотуберкулезным препаратам). Соответственно клинические сведения о пациентах служат источником информации, чтобы разделить последовательности на группы устойчивых и чувствительных к выбранному препарату. Таким образом, проблема поиска значимых мутаций в геномах микроорганизмов может быть определена как задача бинарной классификации по известным бинарным признакам.

2. Методы и алгоритмы

Процедура анализа данных включает несколько последовательных шагов (рис. 1). На начальном этапе анализируются новые геномы микроорганизмов, выполняются первичная предобработка и очистка данных. Следующие шаги ориентированы на анализ популяционной структуры, поиск коррелирующих сайтов, а также генов, находящихся под воздействием положительного отбора.

Рис. 1. Процедура анализа геномных данных

На завершающих шагах оценивается вклад мутаций в развитие лекарственной устойчивости, происходит сравнение и аннотирование результатов.

2.1. Фильтрация и очистка данных

Для уменьшения размерности задачи на этапе предобработки, а также для улучшения качества искомых решений применяется ряд фильтров, позволяющих сократить число столбцов в матрице генотипов. Наиболее широко используемым фильтром является фильтр частоты редкой аллели (minor allele frequency, MAF). Как правило, если мутация (редкая аллель) встречается менее чем в 1 % организмов, то она считается недостоверной и может быть исключена из дальнейшего анализа. В исследуемых данных лишь порядка 30 % мутаций прошли фильтр MAF.

Чтобы дополнительно уменьшить количество параметров в моделях, была выполнена процедура удаления дубликатов из матрицы генотипов X. Для этого столбцы матрицы X были разбиты на группы, состоящие из абсолютно идентичных друг другу столбцов; в конечной же матрице был оставлен один представитель из каждой группы. Если в последующих тестах позиция, соответствующая какому-либо столбцу из выбранной группы, признавалась значимой, то в результирующий список включались все позиции, соответствующие остальным столбцам этой группы.

По итогам первичной предобработки данных была получена матрица генотипов, каждый столбец которой уникален и содержит не менее 1 % различий. После объединения исходных VCF-файлов, их фильтрации по качеству, критерию MAF и удаления дубликатов удалось сократить число тестируемых генетических маркеров с 50 000 в сырых данных до порядка 1000 в обработанной информации.

Отфильтрованные и очищенные данные были сгруппированы в наборы, на основе которых происходило последующее обучение. Каждый набор данных представляет собой матрицу, составленную из части строк матрицы генотипов X и соответствующих этим строкам элементов вектора фенотипов Y. Выбор строк осуществлялся в зависимости от тестируемого препарата и используемых условий (например, исследовались генетические маркеры лекарственной устойчивости к офлоксацину при допущении, что все образцы текущего набора обладали чувствительностью к инъекционным препаратам второго ряда).

2.2. Исследование популяционной структуры

Основной целью анализа популяционной структуры является установка принадлежности исследуемого организма к уже известным таксономическим группам и проведение кластеризации данных. Как правило, для определения групповой структуры генетических последовательностей используются методы филогенетического анализа, сполиготипирования и метод главных компонент.

В данной работе для уменьшения размерности, группирования данных и расчета поправок для последующих статистических тестов ассоциаций использовался метод главных компонент [2]. Результаты сполиготипирования образцов из полученных групп применялись для определения принадлежности каждой группы к одному из известных семейств M. tuberculosis. В качестве входящих данных использовалась матрица генотипов X с элементами $x_{ij} \in \{0, 1\}$, строки которой $i = \overline{1,n}$ соответствуют организмам, а столбцы $j = \overline{1,m}$ – упорядоченным позициям (сайтам) выравнивания, содержащим мутации. Элементы матрицы X предварительно нормировались:

$$\bar{x}_{ij} = (x_{ij} - \hat{x}_j) \Big/ \sqrt{\hat{x}_j (1 - \hat{x}_j)},$$

где $\hat{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$ – выборочное среднее для j-го столбца.

По нормированной матрице $\bar{X}$ вычислялась ковариационная матрица $C = \frac{1}{n-1} \bar{X} \bar{X}^T$. Согласно сингулярному разложению нормированная матрица генотипов представима в виде $\bar{X} = USV^T$. Тогда с точностью до константы выполнено $C = US^2U^T$, где U представляет собой матрицу $n \times n$, составленную из собственных векторов матрицы $\bar{X}\bar{X}^T$, которые образуют ортонормированный базис. Каждый элемент u_{li} матрицы U может быть интерпретирован как «координата» i-го организма на l-й оси вариаций (главной компоненте). Для визуализации использовались проекции на первые две главные компоненты (рис. 2), которые отражают популяционную структуру, где генетически наиболее близкие друг другу организмы объединены в группы.

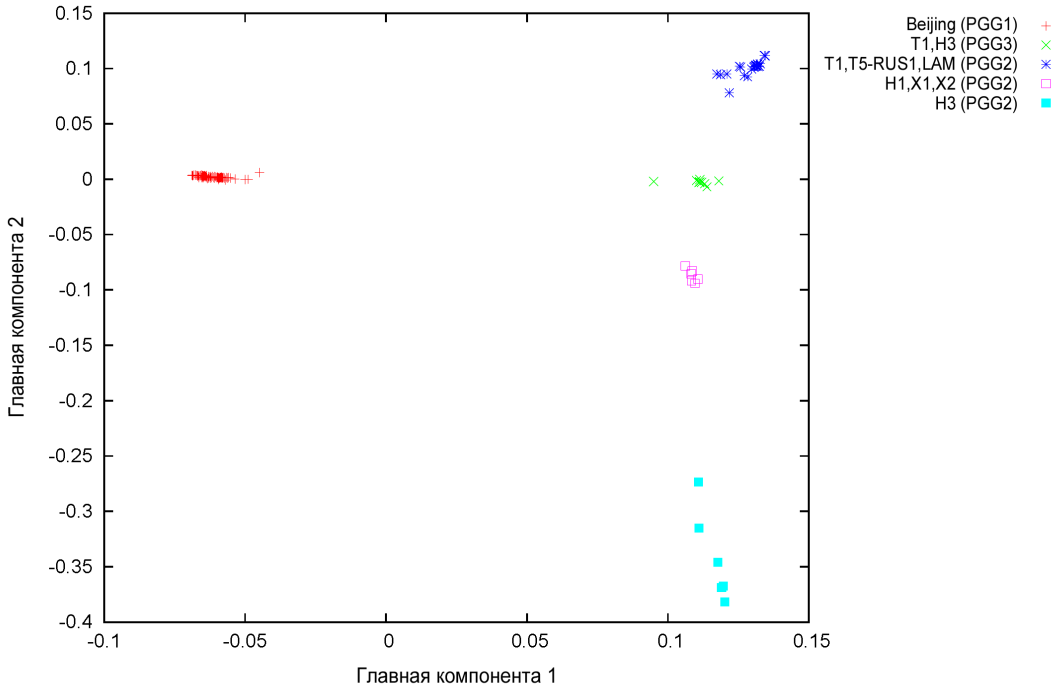

Рис. 2. Расположение популяций M. tuberculosis, представленных в Беларуси, в плоскости двух главных компонент

На полном наборе данных, состоящем из 132 организмов и 8000 геномных маркеров, критерий отбора по правилу Кайзера оставляет 28 главных компонент в качестве значимых.

2.3. Одномаркерные методы анализа для поиска ассоциированных мутаций

Методы одномаркерного анализа, как правило, являются модификациями классических тестов статистической проверки гипотез и используются для проверки существования зависимостей между отдельными мутациями и изменением фенотипа. В данном случае исследуется статистическая связь между наличием мутации в некоторой позиции генома и развитием лекарственной устойчивости к рассматриваемому препарату. Для этого составляется таблица сопряженности, где в качестве признаков выбираются наименование препарата (R – устойчивость к этому препарату, S – чувствительность к нему) и мутация (Y – присутствие, N – отсутствие). Подобные таблицы строятся независимо друг от друга для каждой комбинации, состоящей из наименования препарата и мутации. Если рассматриваемая позиция находится в пределах некоторого гена, то при наличии мутации также говорят о присутствии альтернативной аллели гена, а в случае ее отсутствия – преобладающей аллели.

Предположим, что исследуемая группа состоит из n_{0*} лекарственно-устойчивых бактерий и n_{1*} лекарственно-чувствительных (табл. 1). Задача статистической проверки гипотез заключается в формулировании основной нулевой гипотезы H_0 о независимости признаков и последующей попытке ее опровержения в пользу альтернативной гипотезы H_1.

Таблица 1

Таблица сопряженности, рассматриваемая в одномаркерных тестах поиска ассоциаций

Чувствительность к препарату	Мутация присутствует (Y)	Мутация отсутствует (N)	Всего
Устойчив к препарату (R)	n_{00}	n_{01}	n_{0*}
Чувствителен к препарату (S)	n_{10}	n_{11}	n_{1*}
Всего	n_{*0}	n_{*1}	n_{**}

Для проверки статистических гипотез применялись следующие тесты: критерий Кохрана – Мантеля – Хензеля (Cochran – Mantel – Haenszel, CMH) [3] для анализа таблиц сопряженности, построенных по подгруппам организмов, и χ^2-критерий Пирсона с поправкой EIGENSTRAT [4] для учета структуры популяции на основании рассчитанных в подразд. 2.2 значимых главных компонент. Необходимость внесения поправок по отношению к популяционной структуре обусловлена тем, что варьирование частоты встречаемости некоторых мутаций в исследуемых организмах может объясняться их принадлежностью к различным популяциям, а не воздействием лекарственных препаратов.

Статистический тест Кохрана – Мантеля – Хензеля использовался для анализа таблиц сопряженности размерности $2\times 2\times K$, где K означает число популяций. Этот тест позволяет нивелировать влияние фактора стратификации на конечный результат. Нулевая гипотеза H_0 в этом случае заключается в предположении об условной независимости двух признаков (наличие лекарственной устойчивости, тип аллели) в случае принадлежности исследуемых организмов к одной из групп. Вероятностная модель при нулевой гипотезе предполагает, что маргинальные итоги фиксированы для каждой из K групп организмов. Тогда число наблюдений $n_{00(k)}$ в таблице с номером k описывается гипергеометрическим распределением вероятностей $n_{00(k)} \sim HG\left(n_{0*(k)}, n_{**(k)}, n_{*0(k)}\right) = \binom{n_{0*(k)}}{n_{00(k)}}\binom{n_{1*(k)}}{n_{10(k)}} \Big/ \binom{n_{**(k)}}{n_{*0(k)}}$, где $\binom{n}{p} = \dfrac{n!}{p!(n-p)!}$ обозначает биномиальный коэффициент. Отсюда можно рассчитать значения и в остальных ячейках таблицы. Если учесть, что математическое ожидание и дисперсия гипергеометрического распределения вероятностей имеют вид $E\left\{n_{00(k)}\right\} = \dfrac{n_{0*(k)}n_{*0(k)}}{n_{**(k)}}$ и $Var\left\{n_{00(k)}\right\} = \dfrac{n_{0*(k)}n_{1*(k)}n_{*0(k)}n_{*1(k)}}{n_{**(k)}^2\left(n_{**(k)}-1\right)}$ соответственно, то статистика Кохрана – Мантеля – Хензеля задается формулой

$$\text{CMH} = \frac{\left|\sum_{k=1}^{K}\left(n_{00(k)} - E\left\{n_{00(k)}\right\}\right)\right|^2}{\sum_{k=1}^{K} Var\left\{n_{00(k)}\right\}} \qquad (1)$$

и имеет распределение χ_1^2 при $n\to\infty$ [5].

Выполнение гипотезы H_0 также эквивалентно равенству единице отношений шансов $R_{(k)} = \dfrac{n_{00(k)}/n_{10(k)}}{n_{01(k)}/n_{11(k)}}$, $k=\overline{1,K}$, в каждой из K подгрупп, т. е. H_0: $R_{(1)} = R_{(2)} = ... = R_{(k)} = 1$. Следует отметить, что тест валиден, когда отношения шансов имеют одинаковые направления и масштабы, и не является эффективным для случая, когда существуют подгруппы K_1 и K_2, такие, что для любого элемента $k_1 \in K_1$ отношение шансов $R_{(k_1)} > 1$ и в то же время для любого элемента $k_2 \in K_2$ отношение шансов $R_{(k_2)} < 1$. В численных экспериментах на предоставленных данных использовалась версия теста, реализованная в пакете PLINK [6].

Еще одним способом учесть зависимости между наблюдениями ввиду генетического сходства соответствующих им организмов, принадлежащих одной популяционной группе, является внесение поправок в данные, которые поступают на вход классических статистических тестов. В частности, использовался классический χ^2-критерий Пирсона с поправкой на популяционную структуру, реализованный в пакете EIGENSTRAT [4]. Эта поправка представляет собой модификацию матрицы генотипов X, по которой затем вычисляется тестовая статистика. Пусть x_{ij}, $i=\overline{1,n}$, $j=\overline{1,m}$ – элемент исходной матрицы X, где i – номер организма, j – позиция мутации. Будем считать, что ее элементы были предварительно нормированы описан-

ным в подразд. 2.2 способом. Пусть ее сингулярное разложение имеет вид $X = USV^T$. Тогда для каждой l-й оси вариаций элементы скорректированной матрицы генотипов рассчитываются по формуле $x_{ij(l)}^{adj} = x_{ij} - \gamma_j u_{li}$, где $\gamma_j = \sum_{i=1}^{n} u_{li} x_{ij} \Big/ \sum_{i=1}^{n} u_{li}^2 = \sum_{i=1}^{n} u_{li} x_{ij}$. В матричном виде эти соотношения записываются как $X_l^{adj} = X - Xu_l$, т. е. из данных вычитается проекция на l-ю главную компоненту. Предполагается, что данные были предварительно центрированы относительно начала координат, а величина проекции расстояния между любыми двумя наблюдениями вдоль выбранной главной компоненты пропорциональна генетическому расстоянию между соответствующими организмами. Таким образом устраняются значительные вариации, вызванные принадлежностью организмов к различным популяциям. В силу того что оси главных компонент ортогональны, эта поправка может быть вычислена независимо для каждой оси. Поправка для вектора фенотипов рассчитывается как $y_{(l)}^{adj} = y_l - u_l$.

В результате статистической проверки гипотез во всех представленных тестах вычисляется значение некоторой статистики, имеющей известное распределение, и рассчитывается *p*-значение. Общепринятым значением ошибки первого рода в полногеномных тестах ассоциаций считается $\alpha = 5 \times 10^{-8}$ [7]. В случае применения стандартного порога $\alpha = 0{,}01$ для ошибки первого рода вероятность сделать хотя бы одну ошибку уже при 100 тестах составляет $1 - \left(1 - 10^{-2}\right)^{100} \approx 0{,}63$, а при 1000 тестах – $1 - \left(1 - 10^{-2}\right)^{1000} \approx 1$. Для уменьшения вероятности ошибок при проведении множественной проверки гипотез используются методы поправки *p*-значений: поправки Бонферони, Бонферони – Холма [8] и Бенджамини – Хохберга [9].

При представлении результатов в работе использовалась поправка Бенджамини – Хохберга. Она гарантирует, что ожидаемая доля ложных отклонений гипотез (false discovery rate, FDR) не будет превосходить α. Алгоритм проверки гипотез по данному методу работает итерационно с упорядоченным по возрастанию набором *p*-значений, отвергая все гипотезы, для которых $p_{(j)} < \alpha \cdot j / n$.

2.4. Многомаркерные методы анализа для поиска ассоциированных мутаций

Несмотря на применяемые поправки, большим недостатком одномаркерных методов является то, что они не учитывают парные взаимодействия и взаимодействия более высоких порядков между исследуемыми генетическими вариациями. В то же время существуют методы, именуемые в литературе многомаркерными методами анализа ассоциаций, которые могут реалистично моделировать влияние множества генетических маркеров одновременно. Однако применение многомаркерного анализа сопряжено с трудностями, возникающими при решении задач большой размерности, когда число исследуемых мутаций значительно превосходит доступное число наблюдений. Чтобы сгладить влияние этих факторов, применялись методы отбора признаков и способы регуляризации классических методов машинного обучения.

2.4.1. Методы на основе регуляризованной логистической регрессии

Логистическая регрессия с регуляризацией вектора параметров по l_1-норме (метод Лассо) в выполненных экспериментах показала лучшие результаты по сравнению с регуляризацией по l_2-норме (гребневая регрессия). Предпочтительным вариантом стало применение метода Elastic Net (компромиссный вариант между регуляризацией по l_1- и l_2-нормам). В работе [10] доказывается, что при регуляризации Лассо для корректной оценки параметров достаточно иметь выборку, число наблюдений в которой оценивается как логарифм от числа параметров модели. Формально модель логистической регрессии задается соотношением $p(x) = 1 \Big/ \left(1 + e^{-\left(\beta_0 + x\boldsymbol{\beta}^T\right)}\right)$, где $p(x)$ соответствует вероятности принадлежности генотипа x к классу лекарственно-устойчивых бактерий, $\boldsymbol{\beta}$ – вектор коэффициентов.

Обозначим через $l(x_i, y_i, \boldsymbol{\theta})$ функцию потерь при классификации с помощью логистической регрессии i-го организма с вектором генотипа x_i, значением фенотипа y_i и вектором параметров $\theta = (\beta_0, \boldsymbol{\beta})$ размерности m+1. Тогда для поиска оценок вектора параметров минимизируется функционал

$$\sum_{i=1}^{n} l(x_i, y_i, \boldsymbol{\theta}) + \lambda_1 \|\boldsymbol{\theta}\|_{l_1} + \lambda_2 \|\boldsymbol{\theta}\|_{l_2} = \sum_{i=1}^{n} l(x_i, y_i, \boldsymbol{\beta}) + \lambda \left(\alpha \sum_{j=0}^{m} |\beta_j| + (1-\alpha) \sum_{j=0}^{m} \beta_j^2 \right) \to \min_{\boldsymbol{\beta}}.$$

Для решения данной оптимизационной задачи использовался метод циклического покоординатного спуска [11], реализованный в библиотеке LARS пакета R. Несмотря на то что целевая функция при такой регуляризации не является дифференцируемой из-за недифференцируемости модуля, она обладает полезным свойством разреженности, т. е. обращением в нуль коэффициентов при увеличении параметра λ_1. Для признания коэффициентов значимыми при некотором λ и $\alpha \approx 0{,}7$, которые были выбраны как наиболее предпочтительные при обучении на имеющихся наборах данных, использовалось ограничение на малость коэффициентов, т. е. значимыми признавались мутации из подмножества S, такие, что для любого $s \in S$ абсолютное значение коэффициента $|\beta_s| \geq \tau$. Одним из главных недостатков подобной регуляризации является включение в итоговый результат шумовых признаков до того, как были включены все значимые признаки, и ошибки при отборе коррелирующих признаков. Подход с использованием сетей релевантности признаков, описанный в подразд. 2.5, позволяет частично корректировать найденные решения и получать более точный результат даже в случае переобучения.

2.4.2. Линейная смешанная модель

Линейная смешанная модель в статистической генетике применяется для анализа геномных последовательностей, для которых характерна сложная парадигма наследственности и изменчивости. Наиболее широкое распространение в задачах вычислительной биологии получила так называемая смешанная линейная модель, описанная Хендерсоном [12]. В общем виде смешанная модель может быть представлена в матричной форме:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{2}$$

где $\boldsymbol{y}$ – вектор фенотипов; X – матрица генотипов; $\boldsymbol{\beta}$ – вектор параметров, определяющих значимости мутаций для развития лекарственной устойчивости; $\boldsymbol{u}$ – вектор случайных эффектов, имеющий многомерное нормальное распределение $\boldsymbol{u} \sim N_n(0, \lambda\sigma_\varepsilon^2 K)$. Здесь $\lambda = \sigma_u^2 / \sigma_\varepsilon^2$ – отношение между дисперсиями, обусловленными внутренними (генетическими) и внешними факторами, σ_ε^2 – дисперсия случайной ошибки, а K – матрица родства организмов (kinship matrix) размерности $n \times n$, где n – число исследуемых организмов. Таким образом, благодаря добавлению случайных эффектов модель позволяет учитывать следствия от наличия популяционной структуры. Одним из способов задания матрицы родства является оценка $K = XX^T/m$, где m – число признаков (генетических маркеров). Как правило, матрица K центрируется так, чтобы среднее значение в каждом столбце равнялось нулю. Вектор случайных ошибок $\boldsymbol{\varepsilon}$, называемый также вектором эффектов окружающей среды, имеет многомерное нормальное распределение вида $N_n(0, \sigma_\varepsilon^2 I)$, где I – единичная матрица соответствующей размерности.

Вычисление точечных оценок параметров $\boldsymbol{\beta}$, σ_ε^2 и λ выполняется по методу максимального правдоподобия [13]. Функция правдоподобия для линейной смешанной модели (2) имеет вид $l(\boldsymbol{y}; \lambda, \sigma_\varepsilon, \boldsymbol{\beta}) = -\frac{1}{2}\left(n\log(2\pi\sigma_\varepsilon^2) + \log|H| + \frac{1}{\sigma_\varepsilon^2}(y - X\boldsymbol{\beta})^T H^{-1}(y - X\boldsymbol{\beta}) \right)$, где $H = \lambda K + I$. Если считать параметр λ известным, то максимум функции достигается при оценках параметров

$\hat{\boldsymbol{\beta}} = (X^T H^{-1} X)^{-1} X^T H^{-1} \boldsymbol{y}$ и $\hat{\sigma}_\varepsilon^2 = \left(y - X\hat{\boldsymbol{\beta}}\right)^T H^{-1} \left(y - X\hat{\boldsymbol{\beta}}\right) \Big/ n$. Подставляя эти значения в выражение для функции правдоподобия, получим, что окончательное вычисление оценки максимума правдоподобия сводится к оптимизации по параметру λ функционала

$$l(\boldsymbol{y};\lambda) = -\frac{1}{2}\left(n\log\left(\frac{2\pi\left(y - X\boldsymbol{\beta}\right)^T H^{-1}\left(y - X\boldsymbol{\beta}\right)}{n} \right) + \log|H| + n \right).$$

Для проверки ограничений на параметры модели используется один из классических статистических тестов: тест отношения правдоподобия, тест Вальда или тест множителей Лагранжа. При оценивании значимости отдельных генетических маркеров с индексами $j = \overline{1,m}$ по выборочным данным проверяется нулевая гипотеза $H_0: \beta_j = 0$ против альтернативы $H_1: \beta_j \neq 0$. В частности, в тесте отношения правдоподобия сравниваются оценки максимума функции правдоподобия, полученные в нулевой и альтернативной моделях. Построенная на их основе тестовая статистика $z_{LR} = 2\left(l_1(\mathbf{y};\hat{\lambda}_1) - l_0(\mathbf{y};\hat{\lambda}_0)\right)$ асимптотически имеет распределение χ_1^2, если верна нулевая гипотеза.

На практике представляет интерес оценка степени наследуемости признака (доля дисперсии фенотипа, объясненная уравнением регрессии в нулевой модели, т. е. при $\beta = 0$). Значение данной величины определяется по формуле $PVE = \sigma_u^2 \big/ \left(\sigma_u^2 + \sigma_\varepsilon^2\right) = \lambda \big/ (\lambda + 1)$ и служит аппроксимацией для оценки вклада генетических факторов в изменение фенотипа. В случае когда величина наследуемости близка к единице, изменчивость наблюдаемого фенотипа (например, устойчивость/чувствительность к исследуемому лекарственному препарату) может быть полностью объяснена унаследованными генетическими факторами. Если же величина наследуемости близка к нулю, изменчивость фенотипа в большей степени определяется факторами внешней среды.

2.4.3. Методы стохастического отбора значимых признаков

Алгоритм стохастического поиска важных признаков основан на байесовской процедуре отбора переменных. Он идентифицирует комбинации полиморфизмов (мутаций), которые обладают наилучшей прогностической способностью, выполняя поиск среди моделей линейной регрессии с числом факторов, не превосходящим некоторой величины k, либо иерархических лог-линейных моделей с аналогичным ограничением по числу переменных. Так как ожидаемое число значимых мутаций не является большим, в работе использовались значения $k \in [2,\ 5]$.

Согласно принципу Байеса процедура выбора оптимальной модели состоит в поиске варианта, имеющего максимальную апостериорную вероятность среди моделей рассматриваемого класса. Пусть рассматриваются альтернативные модели $M_1, \ldots, M_L$, образующие множество M, и заданы их априорные вероятности $P(M_i)$. Каждой модели M_i соответствует определенная плотность распределения $p_i(\boldsymbol{\theta})$ и условная плотность $p_i\left(X, \boldsymbol{y} \middle| \boldsymbol{\theta}\right)$, где $\boldsymbol{\theta} \in \Theta_i$ – вектор параметров из множества настраиваемых параметров модели M_i; X – матрица генотипов; $\boldsymbol{y}$ – вектор фенотипов. Тогда апостериорные вероятности моделей M_i при известных наблюдениях X, $\boldsymbol{y}$ вычисляются с помощью формулы Байеса $P\left(M_i \middle| X, \boldsymbol{y}\right) = P\left(X, \boldsymbol{y} \middle| M_i\right) P(M_i) \Big/ \Big/ \sum_{j=1}^{L} P\left(X, \boldsymbol{y} \middle| M_j\right) P\left(M_j\right)$ и для поиска наилучшей модели требуется максимизировать по $\mu \in M$ рассчитанную таким образом функцию апостериорной вероятности $P(\mu \mid X, \boldsymbol{y})$.

Каким бы ни было выбрано множество моделей M, в котором осуществляется поиск, на нем должна быть определена функция $nbh(\mu)$ для всех $\mu \in M$, возвращающая окружение модели μ, т. е. множество предшествующих и последующих моделей. Любые две модели из множества M, например μ и μ', должны быть соединены как минимум одной простой цепью $\mu = \mu_1, \mu_2, ..., \mu_l = \mu'$, такой, что $\mu_j \in nbh(\mu_{j-1})$, $j = \overline{2,l}$.

Алгоритм, непосредственно выполняющий поиск наилучшей модели, известен в литературе как ориентированный на моду стохастический поиск (mode-oriented stochastic search, MOSS) [14], который формирует множество моделей $M(c)$ вида

$$M(c) = \left\{ \mu \in M : P(\mu | X, \boldsymbol{y}) \geq c \cdot \max_{\mu' \in M} P(\mu' | X, \boldsymbol{y}) \right\},$$

где $c \in (0,1)$. В процессе своей работы алгоритм поддерживает множество S текущих моделей, обновляемое по мере поиска. Под $S(c)$ будем понимать подмножество множества S, введенное аналогично тому, как вводится $M(c)$ относительно M. Пусть помимо c задана константа c', такая, что $0 < c' < c$ и $S(c) \subset S(c')$, и предопределена вероятность q удаления моделей из множества $S \setminus S(c)$. Некоторую модель μ будем называть исследованной, если было просмотрено все ее окружение $\mu' \in nbh(\mu)$. Таким образом, множество S содержит исследованные и неисследованные модели. Алгоритм MOSS работает следующим образом:

1. Множество S произвольным образом инициализируется моделями из множества M. Для каждой модели $\mu \in S$ вычисляется и запоминается ее апостериорная вероятность $P(\mu | X, \boldsymbol{y})$, а μ помечается как неисследованная.

2. Пусть $\Lambda \subset S$ – подмножество неисследованных моделей из S. Выберем некоторую модель $\mu \in \Lambda$ случайным образом с вероятностью, пропорциональной апостериорной вероятности $P(\mu | X, \boldsymbol{y})$ модели. Пометим выбранную модель μ как исследованную.

3. Для всех моделей $\mu' \in nbh(\mu)$ проверяем их принадлежность к S. Если $\mu' \notin S$, то вычисляем и запоминаем ее апостериорную вероятность $P(\mu' | X, \boldsymbol{y})$. Если при этом $\mu' \in S(c')$, то добавляем μ' в S и помечаем ее как неисследованную. Если теперь μ' имеет наибольшую апостериорную вероятность среди всех моделей в S, то из S удаляются модели $S \setminus S(c')$.

4. С вероятностью q из S удаляются все модели $S \setminus S(c)$.

5. Если все модели в S оказываются исследованными, то из S удаляются модели $S \setminus S(c)$ и множество $S(c)$ является ответом. Иначе возвращаемся к п. 2 алгоритма.

На примере алгоритма MOSS видно, что модели с низкой апостериорной вероятностью автоматически удаляются из рассмотрения, а приоритет на каждой итерации поиска отдается исследованию наиболее перспективных моделей среди просмотренных ранее, что позволяет алгоритму быстро двигаться в сторону моделей с высокой апостериорной вероятностью. Значения гиперпараметров c, c' и q определяют размер множества просматриваемых моделей, что позволяет ограничивать их перебор.

В отличие от регрессионного подхода к оценке значимости индивидуальных мутаций алгоритм MOSS позволяет ранжировать и сравнивать модели в соответствии с их апостериорной вероятностью. Кроме того, апостериорные вероятности моделей позволяют оценивать значимость отдельных мутаций с помощью байесовского усреднения [15].

2.5. Использование графической модели для отбора значимых признаков

Важно отметить, что описанные выше методы, за исключением в какой-то степени алгоритма MOSS, не учитывают взаимозависимости между признаками исследуемой модели. В реальных же данных генетические маркеры достаточно часто коррелируют друг с другом

с коэффициентом корреляции Пирсона, превосходящим по модулю 0,8. Одним из способов представления пространства признаков, который учитывает их взаимозависимости, является так называемая сеть релевантности признаков [16].

Рассмотрим задачу построения сети релевантности признаков для поиска значимых мутаций в некоторой геномной последовательности. Допустим, что поставлена задача обучения с учителем по m признакам с размером обучающей выборки n. Тогда сетью релевантности признаков называют бинарное марковское случайное поле, заданное с помощью случайного вектора $\Xi=(\Xi_1,\Xi_2,...,\Xi_m)\in\{0,1\}^m$ и неориентированного графа $G(V,E)$. Состояние вершины $j\in V$ графа $G(V,E)$ кодируется при помощи бинарной переменной ξ_j, принимающей значение $\xi_j=1$, если наличие j-й мутации связано с развитием лекарственной устойчивости, и $\xi_j=0$, если такая связь не установлена. Любые вершины i и j, соответствующие попарно коррелированным мутациям, соединены в графе неориентированным ребром $(i,j)\in E$. На вершинах графа $G(V,E)$ введем унарные потенциалы $\phi(\Xi_j)$, характеризующие априорную вероятность того, насколько j-я мутация определяет лекарственную устойчивость к рассматриваемому лекарственному препарату, а на ребрах – парные потенциалы $\psi(\Xi_i,\Xi_j)$. Для заданной таким образом сети вероятность некоторой реализации $\boldsymbol{\xi}=(\xi_1,\xi_2,...,\xi_m)$ случайного вектора Ξ определяется формулой

$$P(\boldsymbol{\xi})=\frac{1}{Z}\prod_{j=1}^{|V|}\phi_j(\xi_j)\prod_{(i,j)\in E}\psi_{ij}(\xi_i,\xi_j)=\frac{1}{Z}\exp\left(\sum_{j=1}^{|V|}\log\phi_j(\xi_j)+\sum_{(i,j)\in E}\log\psi_{ij}(\xi_i,\xi_j)\right), \quad (3)$$

где Z – константа нормализации, а $|V|=m$. Определим функцию энергии сети $E(\boldsymbol{\xi})$ по формуле

$$E(\boldsymbol{\xi})=\sum_{j=1}^{|V|}E_j(\xi_j)+\sum_{(i,j)\in E}E_{ij}(\xi_i,\xi_j)+E_0, \quad (4)$$

где $E_j(\xi_j)=-\log\phi(\xi_j)$, $E_{ij}(\xi_i,\xi_j)=-\log\psi(\xi_i,\xi_j)$, E_0 – константа. Тогда конечная цель заключается в поиске конфигурации $\boldsymbol{\xi}_{opt}=(\xi_1,\xi_2,...,\xi_m)$, которая минимизирует функцию энергии $E(\boldsymbol{\xi})$ в формуле (4), что эквивалентно максимизации вероятности $P(\boldsymbol{\xi})$ в формуле (3). Скрининг переменных с помощью сетей релевантности состоит из двух этапов: построения сети и статистического вывода в ней.

На этапе построения сети требуется задать функции ϕ и ψ. Чтобы задача имела решение за полиномиальное время, выбор функции ψ выполняется с учетом условия субмодулярности [16]:

$$E_{ij}(0,\ 0)+E_{ij}(1,\ 1)\le E_{ij}(0,\ 1)+E_{ij}(1,\ 0), \quad (5)$$

которое является необходимым и достаточным для сведения задачи минимизации функции энергии (4) к задаче поиска минимального разреза в соответствующем ей графе $G'(V',E')$ (способ построения этого графа приведен в конце раздела).

Для обеспечения условия (5) в рассматриваемой задаче можно положить $\psi(\Xi_i,\Xi_j)=\exp\{\lambda\cdot|r_{ij}|\cdot I(\Xi_i=\Xi_j)\}$, где $\lambda>0$ – положительный параметр; r_{ij} – коэффициент корреляции признаков; $I(\Xi_i=\Xi_j)$ – индикаторная переменная, равная единице, если значения

признаков Ξ_i и Ξ_j совпадают. Заметим, что такой выбор функции парных потенциалов поощряет ситуации, когда хорошо коррелированные признаки должны оказаться в одном и том же состоянии, т. е. коррелирующие друг с другом мутации одновременно либо являются, либо не являются маркерами лекарственной устойчивости к рассматриваемому препарату.

Выбор унарных потенциалов вершин ϕ достаточно произволен и можно положить $\phi(\Xi_j) = \exp\{|\Xi_j - q_j|\}$, где $q_j = 1 - p_j$ – вероятность того, что j-й признак не является существенным. Предположим, что при определении значимости каждого отдельного признака использовались методы статистической проверки гипотез, например методы из подразд. 2.3, 2.4, и для каждого признака было вычислено значение некоторой статистики. Пусть $\boldsymbol{T}$ – вектор этих статистик $\boldsymbol{T} = (T_1, \ldots, T_m)$, где все T_j независимы. Тогда каждое p_j определяется исходя из значений тестовой статистики T_i. Например, $p_j = 1$ при $|T_j| \geq \tau$ и $p_j = 0$ в противном случае.

Вторым шагом при определении значимости признаков является вероятностный вывод. Для построенной сети релевантности признаков требуется найти вектор ξ_{opt}, максимизирующий апостериорную вероятность (3). Подставив в формулу (4) выражения для потенциалов вершин и ребер, получим эквивалентную формулировку в виде задачи минимизации функции энергии:

$$\sum_{j=1}^{|V|} |\xi_j - p_j| + \lambda \sum_{i,j=1}^{|V|} I(\xi_i \neq \xi_j) |r_{ij}| \to \min_{\xi}. \qquad (6)$$

Поскольку парные потенциалы $\psi(\Xi_i, \Xi_j)$ подобраны так, что полученные на их основе функции $E_{ij}(\Xi_i, \Xi_j) = \lambda I(\Xi_i \neq \Xi_j) |r_{ij}|$ удовлетворяют условиям субмодулярности (5), наиболее вероятная реализация случайного вектора Ξ может быть найдена с помощью алгоритмов поиска минимального *s*-*t*-разреза в специальном графе $G'(V', E')$, который получается из графа $G(V, E)$ следующим образом. Множество вершин $V' = V \cup \{s, t\}$ нового графа формируется путем введения двух дополнительных терминальных вершин s, t. Для построения множества дуг E' каждая нетерминальная вершина $j \in V$ соединяется с истоком s ребром веса $c_{sj} = |1 - p_j| - |0 - p_j|$, если $|1 - p_j| - |0 - p_j| > 0$, либо со стоком t ребром веса $c_{jt} = |0 - p_j| - |1 - p_j|$ в противном случае. Каждая пара нетерминальных вершин $i, j \in V$ соединяется ребром, вес которого определяется как $c_{ij} = E_{ij}(0, 1) + E_{ij}(1, 0) - E_{ij}(0, 0) - E_{ij}(1, 1)$. Отметим, что благодаря ограничениям (5) и выбору функций потенциалов веса дуг в графе $G'(V', E')$ являются неотрицательными. Любой *s*-*t*-разрез такого графа соответствует разбиению множества вершин V' на подмножества истока и стока. После нахождения минимального *s*-*t*-разреза в графе $G'(V', E')$ значения переменных ξ_j определяются следующим образом: $\xi_j = 1$, если вершина j остается связанной со стоком t в полученном разрезе, и $\xi_j = 0$ в противном случае. Величина разреза, рассчитываемая как сумма емкостей всех его ребер, совпадает со значением функционала энергии (4) при таких значениях переменных ξ.

3. Результаты исследования

По итогам анализа имеющихся данных полногеномного секвенирования M. tuberculosis и медицинских карт пациентов были отобраны 132 генома и сформированы тестовые наборы данных (табл. 2). Материал из каждого набора был проанализирован с использованием описанной выше методологии.

Таблица 2

Характеристики наборов данных, сформированных на основе результатов лабораторных тестов на чувствительность к лекарственным препаратам для проведения вычислительного эксперимента

Набор данных	Число наблюдений	Условия отбора	Лекарственные препараты, к которым выполняется поиск маркеров лекарственной устойчивости
1	132	–	Все первой линии и офлоксацин
2	48	Чувствителен к аминогликозидам	Офлоксацин
3	54	Чувствителен к офлоксацину	Аминогликозиды
4	23	Чувствителен к аминогликозидам, но устойчив к рифампицину	Офлоксацин
5	63	Чувствителен к амикацину	Капреомицин
6	122	–	Аминогликозиды
7	29	Чувствителен к офлоксацину, но устойчив к рифампицину	
8	48	Чувствителен к аминогликозидам	Все второй линии кроме аминогликозидов
9	122	–	

Для сопоставления результатов отбора значимых мутаций с помощью предложенной методологии с уже известными генетическими маркерами лекарственной устойчивости была использована публичная база данных TBDreamDB [17], которая содержит информацию о большинстве описанных в литературе мутаций к препаратам первого ряда, а также списки полиморфизмов, которые заложены в применяемых на практике тест-системах GenoType MTBDRplusV2 и GenoType MTBDRsl [18, 19].

Для визуализации результатов полногеномных тестов ассоциаций применяют так называемые Манхэттен-графики, в которых ось x соответствует позициям однонуклеотидных полиморфизмов в геноме, а ось y – отрицательному десятичному логарифму p-значений (рис. 3). Чем выше статистическая значимость отдельной мутации, тем больше значение координаты y соответствующей точки на графике.

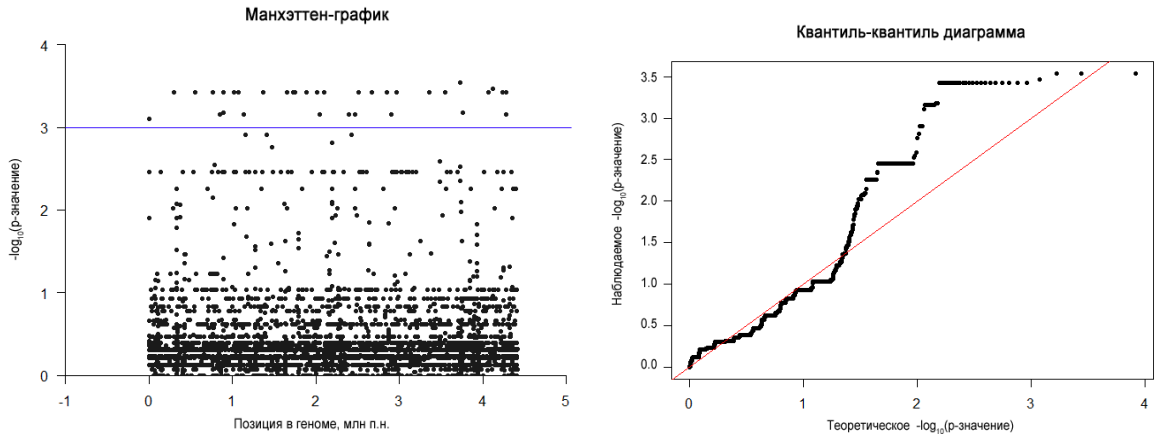

Рис. 3. Манхэттен-график и соответствующая квантиль-квантиль-диаграмма, демонстрирующие результаты применения теста Кохрана – Мантеля – Хензеля для поиска мутаций лекарственной устойчивости к офлоксацину. Для поправки p -значений применялся метод геномной коррекции [20]

Методы одномаркерного анализа ассоциаций, описанные в разд. 2.3, позволили идентифицировать в качестве валидных для определения лекарственной устойчивости как уже известные мутации, так и ряд новых потенциально значимых полиморфизмов. Для контроля качества результатов одномаркерных методов использовались квантиль-квантиль-диаграммы (Q-Q plots) распределений *p*-значений, рассчитанных с учетом поправок на множественные сравнения.

Совокупно по результатам применения смешанных линейных моделей в каждом из экспериментов были найдены множества значимых мутаций и получены оценки параметров модели. Статистическая значимость коэффициентов β оценивалась с помощью теста отношения правдоподобия. При обучении модели применялся скользящий контроль по пяти блокам (5-fold cross-validation) и вычислялись метрики, общепринятые для оценки качества классификации: точность (precision), полнота (recall), F-мера (среднее гармоническое между точностью и полнотой), правильность (accuracy). Истинные значения классов для каждого наблюдения (генома) заимствуются из результатов лабораторных тестов на чувствительность к препарату (табл. 3).

Таблица 3

Результаты классификации для тривиального решающего правила $\alpha(X,\beta) = \text{sign}(X\beta)$, где вектор параметров β оценивался с помощью линейной смешанной модели. Значения $\alpha(X,\beta) = 1$ соответствуют классу $y = 1$ (устойчивость), значения $\alpha(X,\beta) = -1$ – классу $y = 0$ (чувствительность)

Набор данных	Препарат	1	0	Значение истинное/предсказанное				Точность	Полнота	F1	Правильность
				1/1	1/0	0/1	0/0				
1	OFLO	69	63	0,523	0,000	0,417	0,061	0,557	1,000	0,715	0,583
	EMB	102	30	0,083	0,689	0,000	0,227	1,000	0,108	0,195	0,311
	INH	106	26	0,803	0,000	0,000	0,197	1,000	1,000	1,000	1,000
	PZA	28	6	0,824	0,000	0,029	0,147	0,966	1,000	0,983	0,971
	RIF	106	26	0,803	0,000	0,000	0,197	1,000	1,000	1,000	1,000
	STM	110	22	0,826	0,008	0,000	0,167	1,000	0,991	0,995	0,992
2	OFLO	10	38	0,208	0,000	0,000	0,792	1,000	1,000	1,000	1,000
3	AMIK	8	46	0,148	0,000	0,482	0,370	0,235	1,000	0,381	0,519
	CAPR	10	40	0,200	0,000	0,660	0,140	0,233	1,000	0,377	0,340
4	OFLO	10	13	0,435	0,000	0,000	0,565	1,000	1,000	1,000	1,000
6	AMIK	59	63	0,484	0,000	0,516	0,000	0,484	1,000	0,652	0,484
	CAPR	66	51	0,564	0,000	0,436	0,000	0,564	1,000	0,721	0,564
9	CYCL	46	70	0,397	0,000	0,603	0,000	0,397	1,000	0,568	0,397
	ETH	33	89	0,271	0,000	0,631	0,098	0,300	1,000	0,462	0,369
	PARA	22	100	0,115	0,066	0,066	0,754	0,636	0,636	0,636	0,869

Подробнее поясним содержащуюся в табл. 3 информацию. Для примера рассмотрим первую строку этой таблицы. В столбце «Набор данных» приведен номер набора, из которого взяты данные для анализа. В данном случае согласно условиям отбора анализировались все имеющиеся последовательности (см. табл. 2). Название препарата OFLO говорит о том, что выполнялся поиск геномных маркеров лекарственной устойчивости к офлоксацину. Следующие два столбца информируют о составе обучающей выборки: присутствовало 69 образцов, устойчивых к офлоксацину, и 63 чувствительных к нему. Оставшиеся столбцы характеризуют качество предсказаний по набору выявленных методом значимых мутаций: 0,523 – доля образцов в общей выборке, которые являются устойчивыми к офлоксацину и были предсказаны как устойчивые; 0,000 – доля образцов, которые являются устойчивыми, но были отнесены к классу чувствительных; 0,417 – доля образцов, которые являются чувствительными, но были предсказаны как устойчивые; 0,061 – доля образцов, которые являются чувствительными и были отнесены к чувствительным; 0,557 – доля истинно устойчивых среди предсказанных как устойчивые; 1,000 – доля предсказанных как устойчивые среди истинно устойчивых; 0,715 – среднее гармоническое по двум предыдущим показателям; 0,583 – доля правильных предсказаний по всей выборке.

По итогам работы алгоритма стохастического отбора признаков MOSS были получены вероятностные оценки значимости не только отдельных мутаций, но и содержащих их моделей. Такой подход позволяет начать отбор значимых признаков с ранжирования моделей. Несмотря на простоту моделей, среди которых осуществляется перебор вариантов, здесь в них учитываются и возможные взаимодействия между рассматриваемыми признаками. Поскольку общее число признаков после фильтрации данных оставалось достаточно большим (порядка 770 полиморфизмов), была использована двухэтапная схема поиска наилучшей модели. На первом этапе проводился перебор моделей на множестве линейных регрессий с числом регрессоров, не превосходящим трех. На втором этапе взятые признаки послужили построению наиболее вероятной лог-линейной модели, учитывающей взаимодействия между генетическими маркерами. Однако фаза отбора признаков, реализованная в библиотеке genMOSS, в случае большого числа моделей с примерно равной апостериорной вероятностью, эквивалентна полному перебору. Чтобы сократить время вычислений, для формирования входного набора признаков использовались результаты линейной смешанной модели и регуляризованной логистической регрессии. Множества найденных с их помощью мутаций, учитывая поправки на корреляционную структуру признаков, значительно меньше исходного множества признаков. Это позволяет увеличивать число регрессоров в рассматриваемых лог-линейных моделях до пяти-шести. Дальнейшее увеличение сложности моделей приводит к возрастанию числа ложноположительных результатов и усложняет биологическую интерпретацию результатов. При обучении моделей применялись методы скользящего контроля и вычислялись классические метрики оценки качества классификации (табл. 4).

Таблица 4

Результаты классификации с помощью лучших лог-линейных моделей, найденных MOSS

Набор данных	Препарат	1	0	Значение истинное/предсказанное				Точность	Полнота	F1	Правильность
				1/1	1/0	0/1	0/0				
1	OFLO	69	63	0,394	0,130	0,030	0,446	0,929	0,752	0,831	0,840
	EMB	102	30	0,759	0,015	0,030	0,197	0,962	0,981	0,972	0,955
	INH	106	26	0,788	0,015	0,000	0,196	1,000	0,981	0,990	0,985
	PZA	28	6	0,813	0,000	0,000	0,187	1,000	1,000	1,000	1,000
	RIF	106	26	0,698	0,105	0,000	0,197	1,000	0,869	0,930	0,895
	STM	110	22	0,795	0,038	0,000	0,167	1,000	0,954	0,977	0,962
2	OFLO	10	38	0,206	0,000	0,022	0,772	0,902	1,000	0,949	0,978
3	AMIK	8	46	0,103	0,034	0,000	0,863	1,000	0,750	0,857	0,966
	CAPR	10	40	0,160	0,040	0,000	0,800	1,000	0,800	0,889	0,960
4	OFLO	10	13	0,386	0,050	0,029	0,536	0,931	0,885	0,908	0,921
6	AMIK	59	63	0,410	0,074	0,000	0,515	1,000	0,847	0,917	0,926
	CAPR	66	51	0,413	0,152	0,000	0,436	1,000	0,731	0,845	0,848
9	CYCL	46	70	0,180	0,217	0,179	0,424	0,502	0,453	0,477	0,604
	ETH	33	89	0,073	0,198	0,008	0,721	0,905	0,270	0,415	0,794
	PARA	22	100	0,082	0,099	0,000	0,819	0,929	0,454	0,610	0,901

Применение логистической регрессии с регуляризацией показало, что наилучшие результаты классификации достигались при использовании компромиссной регуляризации Elastic Net с коэффициентом $\alpha = 0,7$ по l_1-норме. Для отбора переменных было задано ограничение на минимальное значение модуля коэффициентов, признаваемых значимыми, равное 10^{-3}. При обучении модели применялся скользящий контроль по 10 блокам (10-fold cross-validation) и вычислялись классические метрики оценки качества классификации (табл. 5).

Таблица 5

Результаты классификации с помощью логистической регрессии, регуляризованной по методу Elastic Net с параметром $\alpha = 0,7$

Набор данных	Препарат	1	0	Значение истинное/предсказанное				Точность	Полнота	F1	Правильность
				1/1	1/0	0/1	0/0				
1	OFLO	69	63	0,515	0,008	0,015	0,462	0,971	0,986	0,978	0,977
	EMB	102	30	0,773	0,000	0,008	0,220	0,990	1,000	0,995	0,992
	INH	106	26	0,803	0,000	0,000	0,200	1,000	1,000	1,000	1,000
	RIF	106	26	0,803	0,000	0,000	0,200	1,000	1,000	1,000	1,000
	PZA	28	6	0,824	0,000	0,000	0,177	1,000	1,000	1,000	1,000
	STM	110	22	0,833	0,000	0,000	0,167	1,000	1,000	1,000	1,000
2	OFLO	10	38	0,208	0,000	0,000	0,792	1,000	1,000	1,000	1,000
3	AMIK	8	46	0,111	0,037	0,000	0,852	1,000	0,75	0,857	0,963
	CAPR	10	40	0,200	0,000	0,000	0,800	1,000	1,000	1,000	1,000
4	OFLO	10	13	0,435	0,000	0,000	0,565	1,000	1,000	1,000	1,000
6	AMIK	59	63	0,451	0,033	0,000	0,516	1,000	0,932	0,965	0,967
	CAPR	66	51	0,564	0,000	0,000	0,436	1,000	1,000	1,000	1,000
9	CYCL	46	70	0,388	0,009	0,000	0,603	1,000	0,978	0,989	0,991
	ETH	33	89	0,271	0,000	0,000	0,730	1,000	1,000	1,000	1,000
	PARA	22	100	0,131	0,049	0,000	0,820	1,000	0,727	0,842	0,951

Стоит отметить, что среди рассмотренных методов алгоритм MOSS при достаточно хороших показателях качества классификации выдает наименьшее число значимых признаков, среди которых, как правило, присутствуют уже описанные в литературе мутации лекарственной устойчивости. Смешанные линейные модели и регуляризованная логистическая регрессия идентифицируют гораздо большее количество значимых признаков, что может свидетельствовать о переобучении этих моделей и включении в итоговое множество шумовых признаков. Одномаркерные тесты также оставляют достаточно большое число мутаций в качестве связанных с развитием лекарственной устойчивости.

Анализ значимых признаков, получаемых на выходе описанных выше алгоритмов, показал, что без дополнительных корректировок эти множества не обладают свойством непротиворечивости, т. е. существуют такие пары коррелирующих мутаций, среди которых одна признается значимой, а другая нет. С помощью методов отбора переменных, основанных на применении сетей релевантности признаков (см. подразд. 2.6), удается сократить множества значимых мутаций как минимум вдвое для большинства тестов за счет удаления потенциально ложноположительных результатов. На рис. 4 показаны графики изменения размеров множества значимых мутаций в зависимости от изменения параметра λ в функционале (6) для теста на лекарственную устойчивость к офлоксацину и соответствующий график изменения F -меры как единой метрики, характеризующей точность и полноту. В большинстве случаев заметное сокращение множества значимых мутаций приводит к незначительному ухудшению качества классификации.

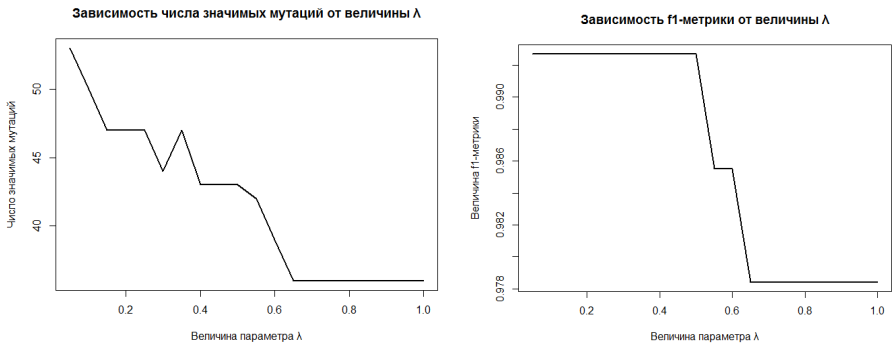

Рис. 4. Графики изменения множества значимых мутаций для лекарственного препарата офлоксацина в результате применения сетей релевантности признаков для корректировки результатов регуляризованной логистической регрессии

При сравнении множества найденных маркеров лекарственной устойчивости с известными ранее были подтверждены многие мутации, входящие в состав широко используемых тест-систем GenoType MTBDRplus/MTBDRsl для препаратов первой линии, а также найдены вероятные мутации, вызывающие резистентность к препаратам второй линии.

Заключение

Поиск по международным базам данных показал, что исследование микобактерии M. tuberculosis является одним из быстро развивающихся направлений. Несмотря на то что в мире неоднократно проводились работы по анализу геномов M. tuberculosis, направленные на поиск мутаций лекарственной устойчивости, такие исследования ограничивались применением методов одномаркерного анализа ассоциаций и изучением препаратов первой линии.

В настоящей работе предложена весьма универсальная методология поиска ассоциаций в геномах патогенных микроорганизмов, позволяющая оценивать вклад мутаций в совокупности и корректировать результаты в случае их противоречивости. Всего в рамках исследования, проводимого в Беларуси, было выполнено секвенирование микобактерий, выделенных у 132 пациентов с различными формами туберкулеза легкого. В дополнение к теоретическому анализу методов осуществлена разработка программного обеспечения в среде R, непосредственно предназначенного для анализа геномов.

Авторы выражают благодарность Республиканскому научно-практическому центру пульмонологии и фтизиатрии Министерства здравоохранения Республики Беларусь за содействие и помощь в предоставлении данных.

Список литературы

1. Borrell, S. Infectiousness, reproductive fitness and evolution of drug-resistant Mycobacterium tuberculosis / S. Borrell, S. Gagneux // Intern. J. Tuberc. Lung Dis. – 2009. – № 13(12). – P. 1456–1466.

2. Patterson, N. Population structure and eigenanalysis / N. Patterson, A. L. Price, D. Reich // PLoS Genet. – 2006. – № 2(12). – P. 2074–2093.

3. Mantel, N. Statistical aspects of the analysis of data from retrospective studies of disease / N. Mantel, W. Haenszel // J. National Cancer Inst. – 1959. – № 22(4). – P. 719–748.

4. Principal components analysis corrects for stratification in genome-wide association studies / A.L. Price [et al.] // Nat. Genet. – 2006. – № 38(8). – P. 904–909.

5. Agresti, A. An introduction to categorical data analysis / A. Agresti. – Wiley, 2002. – Ch. 6. – P. 231–236.

6. PLINK: a tool set for whole-genome association and population-based linkage analyses / S. Purcell [et al.] // Am. J. Hum. Genet. – 2007. – № 81(3). – P. 559–575.

7. Bush, W.S. Genome-wide association studies / W.S. Bush, J.H. Moore // PLoS. Comput. Biol. – 2012. – № 8(12). – P. 1–11.

8. Holm, S. A simple sequentially rejective Bonferroni test procedure testing / S. Holm // Scandinavian Journal of Statistics. – 1979. – № 6. – P. 65–70.

9. Benjamini, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing / Y. Benjamini, Y. Hochberg // Journal of the Royal Statistical Society. – 1995. – № 57. – P. 289–300.

10. Ng, A.Y. Feature selection, L_1 vs. L_2 regularization, and rotational invariance / A.Y. Ng // Proc. of the 21st Intern. Conf. on Machine Learning. – Ban, Canada, 2004.

11. Friedman, J. Regularization paths for generalized linear models via coordinate descent / J. Friedman, T. Hastie, R. Tibshirani // Journal of Statistical Software. – 2010. – № 33(1). – P. 1–22.

12. Henderson, C.R. Sire evaluation and genetic trends / C.R. Henderson // Proc. Anim. Breeding and Genetic Symp. in honor of Dr. J.L. Lush. – Champaign, 1973. – P. 10–41.

13. Zhou, X. Genome-wide efficient mixed-model analysis for association studies / X. Zhou, M. Stephens // Nature Genetics. – 2012. – № 44(7). – P. 821–824.

14. Dobra, A. The Mode oriented stochastic search (MOSS) for log-linear models with conjugate priors / A. Dobra, H. Massam // Statistical Methodology. – 2010. – № 7. – P. 240–253.

15. Applications of the mode oriented stochastic search (MOSS) algorithm for discrete multiway data to genomewide studies / A. Dobra [et al.] // Bayesian Modeling in Bioinformatics. – CRC Press, 2010. – Ch. 3. – P. 63–94.

16. Kolmogorov, V. What energy functions can be minimized via graph cuts? / V. Kolmogorov, R. Zabih // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2004. – № 26(2). – P. 147–159.

17. Tuberculosis drug resistance mutation database / A. Sandgren [et al.] // PLoS Med. – 2009. – № 6(2). – P. 132–136.

18. GenoType MTBDRplus – your test system for a fast and reliable way to detect MDR-TB // Hain Lifesciences [Electronic resource]. – 2015. – Mode of access : http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/genotype-mtbdrplus.html. – Date of access : 10.05.2015.

19. GenoType MTBDRsl – your important assistance for detection of XDR-TB // Hain Lifesciences [Electronic resource]. – 2015. – Mode of access : http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/genotype-mtbdrsl.html. – Date of access : 10.05.2015.

20. Devlin, B. Genomic control for association studies / B. Devlin, K. Roeder // Biometrics. – 1999. – № 55. – P. 997–1004.

Поступила 10.10.2015

[1]*Объединенный институт проблем информатики НАН Беларуси, Минск, Сурганова, 6*
e-mail: roma.sergeev@gmail.com

[2]*EPAM Systems, Минск, Академика Купревича, 1/1*

[3]*Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD, USA*

R.S. Sergeev, I.S. Kavaliou, A.V. Tuzikov, A. Rosenthal, A. Gabrielian

ALGORITHMS FOR IDENTIFYING DRUG-RESISTANCE MUTATIONS IN M. TUBERCULOSIS GENOMES

Analysis of whole-genome sequences often leads to problems of large dimensionality where the number of parameters exceeds the number of available observations. We offer methodology of genome-wide association study and investigate various approaches to assess contribution of mutations in the emergence and development of drug resistance in Mycobacterium tuberculosis. We present the results of our experiments aimed at identifying resistance mutations to the major anti-TB drugs based on data obtained from patients in Belarus.