УДК 004.021

Р. С. СЕРГЕЕВ[1], И. С. КОВАЛЕВ[1,2]

ПОИСК МУТАЦИЙ В ГЕНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ *M. TUBERCULOSIS*, СВЯЗАННЫХ С ЛЕКАРСТВЕННОЙ УСТОЙЧИВОСТЬЮ

[1]Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь, e-mail: roma.sergeev@gmail.com
[2]EPAM Systems, Минск, Беларусь, e-mail: ivanich3d@gmail.com

В настоящей работе исследуются методы полногеномного поиска ассоциаций. Представлена методология, позволяющая осуществлять поиск геномных маркеров лекарственной устойчивости к противомикробным препаратам. Приводятся результаты экспериментов по поиску мутаций резистентности к основным противотуберкулезным препаратам на основании данных о пациентах из Беларуси.

Ключевые слова: полногеномный поиск ассоциаций, биоинформатика, лекарственная устойчивость, регрессия, регуляризация, стохастический поиск, статистическая проверка гипотез.

R. S. SERGEEV[1], I. S. KAVALIOU[1,2]

SEARCH FOR DRUG-RESISTANCE MUTATIONS IN *M. TUBERCULOSIS* GENOME SEQUENCES

[1]United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus, e-mail: roma.sergeev@gmail.com
[2]EPAM Systems, Minsk, Belarus, e-mail: ivanich3d@gmail.com

We investigate algorithms for genome-wide association studies. A methodology is proposed that allows searching for drug-resistance mutations in mycobacterium tuberculosis whole genomes. We present the experimental results on identification of resistance mutations to major anti-TB agents using patients' data from Belarus.

Keywords: genome-wide association, bioinformatics, drug resistance, regression, regularization, stochastic search, statistical hypothesis testing.

Введение. Появление штаммов микроорганизмов, устойчивых к лекарственным препаратам, представляет собой серьезную угрозу здравоохранению. Молекулярные основы резистентности, как правило, связаны с мутационными изменениями в геномах микроорганизмов, возникающими в ходе естественного отбора или под воздействием применяемых препаратов при ненадлежащем лечении. Целью настоящего исследования послужила разработка методологии к анализу мутаций в последовательностях ДНК микобактерий туберкулеза и определение геномных маркеров лекарственной устойчивости, что может иметь большое значение для своевременной диагностики форм туберкулеза и понимания биологических механизмов становления резистентности.

Методика эксперимента. Для выполнения исследования была сформирована выборка из 136 образцов *M. tuberculosis*, включающая 17,7 % штаммов микобактерий туберкулеза, чувствительных ко всем противотуберкулезным препаратам, 2,9 % монорезистентных штаммов и 79,4 % штаммов с множественной лекарственной устойчивостью (МЛУ), среди которых 59,7 % обладали широкой лекарственной устойчивостью (ШЛУ). Все образцы *M. tuberculosis* были получены из клинических проб и секвенированы на платформе Illumina HiSeq2000 с длиной прочтений в 101 пару оснований при среднем покрытии 140×. В ходе сборки геномов короткие

прочтения выравнивались на референсный штамм H37Rv, после чего было выполнено выделение и аннотирование геномных вариаций.

Методология анализа данных. Основываясь на результатах лабораторных тестов, из прочитанных геномов *M. tuberculosis* были сформированы наборы данных для поиска статистически значимых различий между устойчивыми и чувствительными к лекарственным средствам образцами. Процедура анализа данных состояла из нескольких этапов. Первоначально выполнялся сравнительный анализ геномов и исследовалась популяционная структура образцов. Последующие шаги были направлены на изучение зависимостей между геномными полиморфизмами и результатами фенотипических тестов на резистентность. Для этого применялись методы полногеномного поиска ассоциаций, основанные на одно- и многомаркерном анализе зависимостей.

Большинство методов одномаркерного анализа выполняют оценку таблиц сопряженности, в которых приводится информация о частотах основного и альтернативного аллелей в некоторой позиции генома среди устойчивых и чувствительных образцов. Недостатком такого подхода является то, что он исходит из предположения о независимости мутаций друг от друга, что в действительности часто не выполняется. В настоящем исследовании применялся статистический тест Кохрана – Мантеля – Хензеля (Cochran – Mantel – Haenszel, CMH) [1] и χ^2-критерий Пирсона с поправкой Eigenstrat [2], рассчитанной с помощью метода главных компонент. Эти тесты позволяют учесть эффект скрытой переменной, вызванный сегрегацией штаммов по принципу принадлежности к генетически более однородным группам, который называется в популяционной генетике «эффект основателя». Группы выделялись с помощью метода главных компонент, принадлежность входящих штаммов к известным генетическим семействам устанавливалась путем сполиготипирования. При выполнении процедуры множественной проверки статистических гипотез применялась поправка Бенджамини – Хохберга для коррекции *p*-значений.

Более сложные методы анализа зависимостей позволяют явно моделировать совместное влияние геномных маркеров на изменение фенотипа. В настоящей работе использовались регуляризованная логистическая регрессия [3], линейная смешанная модель (linear mixed model LMM) [4, 5] и ориентированный на моду стохастический поиск (mode oriented stochastic search, MOSS) [6]. Приведем краткое описание этих методов.

Предположим, что выборка состоит из n организмов и представлена в виде матрицы генотипов X размера $n \times m$, где m – число исследуемых мутаций. Пусть переменная $y_i \in \{0,1\}$ задает значение фенотипа i-го организма (чувствительный/устойчивый), а $\boldsymbol{x}_i$ – его генотип (строка матрицы X, кодирующая набор исследуемых геномных маркеров). Тогда апостериорная вероятность наблюдения лекарственно-устойчивого фенотипа в классической модели логистической регрессии задается выражением $P(y_i = 1 | \boldsymbol{x}_i) = 1 \Big/ \left(1 + \exp\left(-\boldsymbol{\beta}^T \boldsymbol{x}_i\right)\right)$, где $\boldsymbol{\beta}$ – вектор параметров, определяющих значимость мутаций для развития лекарственной устойчивости. Однако применение логистической регрессии на всех имеющихся полиморфизмах в таком виде сопряжено с рядом трудностей, поскольку число параметров (мутаций) значительно превосходит количество доступных для их оценки наблюдений (секвенированных образцов). С другой стороны, исходя из биологической интерпретации ожидаемого результата, можно предположить, что число значимых маркеров лекарственной устойчивости будет невелико, следовательно, подходящим является разреженное решение. Одним из способов получить такое решение является применение методов регуляризации. Наилучшим способом регуляризации модели логистической регрессии в проведенных экспериментах стало использование метода эластичной сети (elastic net), при котором для поиска оценок вектора параметров минимизировался функционал $\sum_{i=1}^{n} l\left(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}\right) + \lambda_1 \left\|\boldsymbol{\beta}\right\|_{l_1} + \lambda_2 \left\|\boldsymbol{\beta}\right\|_{l_2} -> \min_{\boldsymbol{\beta}}$, где через $l\left(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}\right)$ обозначена функция потерь при классификации i-го организма. Чтобы еще больше сократить число оцениваемых параметров, прогон логистической регрессии был выполнен повторно на множестве геномных маркеров, коэффициенты при которых были признаны статистически значимыми по итогам первого прогона.

В отличие от приведенной выше схемы логистической регрессии, линейная смешанная модель позволяет учесть скрытые взаимосвязи за счет включения в нее вектора случайных эффектов, рассчитанных по матрице родства микроорганизмов. Линейная смешанная модель может быть представлена выражением вида $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{\varepsilon}$, где $\boldsymbol{y}$ – вектор фенотипов, X – матрица генотипов, $\boldsymbol{\beta}$ – вектор параметров. $\boldsymbol{\varepsilon} \sim N_n\left(0, \sigma_\varepsilon^2 I\right)$ – вектор случайных ошибок, σ_ε^2 – дисперсия случайной ошибки, $\boldsymbol{u} \sim N_n\left(0, \sigma_u^2 K\right)$ – вектор случайных эффектов, σ_u^2 – дисперсия случайного эффекта и K – матрица родства организмов (kinship matrix), одним из способов задания которой является оценка $K = XX^T/m$.

Для исследования значимости коэффициентов регрессии в приведенных моделях применялся тест отношения правдоподобия. При этом тестировалась нулевая гипотеза $H_0{:}\beta_j = 0$ против альтернативы $H_1{:}\beta_j \neq 0$ для всех мутаций $j = \overline{1,m}$. Вычисление оценок параметров β_j, σ_ε^2 и σ_u^2 для линейной смешанной модели выполнялось по методу максимального правдоподобия. В случае регуляризованной логистической регрессии для оценки значений $\boldsymbol{\beta}$ и выбора коэффициентов λ_1 и λ_2 применялся алгоритм LARS-EN [3].

Хорошим переборным методом поиска наиболее вероятных моделей регрессии и идентификации значимых полиморфизмов послужил алгоритм ориентированного на моду стохастического поиска на множестве иерархических лог-линейных моделей с числом переменных от двух до пяти [6]. С его помощью удалось проанализировать не только влияние отдельных геномных маркеров, но и их взаимодействие. Согласно алгоритму, для проведения стохастического поиска вводится множество M мощности $|M| = l$, состоящее из рассматриваемых иерархических лог-линейных моделей. Пусть $A_{(k)} \subset \{1,2,...,m\}$, $k = \overline{1,l}$ – некоторое подмножество геномных маркеров, $X_{A_{(k)}}$ – соответствующие им столбцы матрицы генотипов, $\boldsymbol{y}$ – вектор фенотипов, кодирующий результаты тестов на чувствительность к анализируемому препарату. Все модели из множества M можно записать в виде $\mu_k = \{p(X_{A_{(k)}}, \boldsymbol{y} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_k\}, k = \overline{1,l}$, где $\boldsymbol{\theta}$ – вектор параметров и $p(X_{A_{(k)}}, \boldsymbol{y} \mid \boldsymbol{\theta})$ – плотность распределения вероятностей. Зададим также априорные вероятности $P(\mu_k)$ моделей $\mu_k \in M$ и их параметров $p(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_k$. Каждой рассматриваемой иерархической лог-линейной модели ставится в соответствие модель регрессии. Например, для подмножества независимых переменных с индексами из $A = \{i_1, i_2\}$ модель регрессии, порожденная соответствующей иерархической лог-линейной моделью, может быть представлена в виде $\ln \dfrac{P(y_i = 1 \mid X_A)}{P(y_i = 0 \mid X_A)} = \theta(\boldsymbol{y}) + \theta_{i_1}(\boldsymbol{y}, \boldsymbol{x}_{i_1}) + \theta_{i_2}(\boldsymbol{y}, \boldsymbol{x}_{i_2}) + \theta_{i_1 i_2}(\boldsymbol{y}, \boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2})$. Предположим, что имеется алгоритм, с помощью которого можно эффективно вычислять апостериорные вероятности $P(\mu_k \mid X, \boldsymbol{y}) = P(X, \boldsymbol{y} \mid \mu_k) P(\mu_k) / \sum_{t=1}^{l} P(X, \boldsymbol{y} \mid \mu_t) P(\mu_t)$ моделей $\mu_k \in M$ при наличии данных. Будем считать, что априорные вероятности моделей одинаковы, поэтому апостериорная вероятность каждой модели пропорциональна ее маргинальному правдоподобию. В работе [6] показано, что для любого непустого $A \subset \{1,2,...,m\}$ параметры регрессии $y \sim X_A$ и ее маргинальное правдоподобие можно выразить через параметры и маргинальные правдоподобия насыщенных лог-линейных моделей, построенных по множествам переменных $X_A \cup \boldsymbol{y}$ и X_A.

Для каждого элемента множества M введем функцию, возвращающую его окружение, такую что, если μ_k и μ_t – две модели из множества M, соединенные путем $\mu_k = \mu_1, \mu_2, ..., \mu_s = \mu_t$, то μ_i является соседней для модели μ_{i-1} для $i = \overline{2,s}$. В случае иерархических лог-линейных моделей переход от текущего элемента к соседнему может осуществляться одним из следующих способов: включением в модель наименьшего терма, содержащего новый набор взаимодействующих переменных, либо исключением из модели самого длинного терма с выбранными взаимодействующими переменными. Процедура перебора организована таким образом, чтобы при переходах между моделями основное внимание уделять моделям, имеющим наибольшие апостериорные вероятности [6]. В результате своей работы алгоритм формирует множество наиболее подходящих моделей $M(c) = \left\{\mu \in M : P(\mu \mid X, \boldsymbol{y}) \geq c \cdot \max_{\mu' \in M} P(\mu' \mid X, \boldsymbol{y})\right\}$, где $P(\mu \mid X, \boldsymbol{y})$ – апостериорная

вероятность модели $\mu \in M$, а $c \in (0,1)$ – параметр, влияющий на число вариантов для перебора и размер результирующего множества решений.

Рассмотрим все модели $\mu_k \in M(c)$, $k \in B$, где B – множество индексов моделей, вошедших в результирующее множество решений. С помощью процедуры байесовского усреднения по найденным лог-линейным моделям и соответствующим им регрессиям строится классификатор для предсказания результатов теста на лекарственную устойчивость: $P(y \mid X, \boldsymbol{y}) = \sum_{k \in B} P(y \mid \mu_k, X, \boldsymbol{y})\, P(\mu_k \mid X, \boldsymbol{y})$, где $P(y \mid X, \boldsymbol{y})$ соответствует регрессии $y \sim X$, $P(y \mid \mu_k, X, \boldsymbol{y})$ – регрессиям $y \sim X_{A_{(k)}}$ и $P(\mu_k \mid X, \boldsymbol{y})$ – апостериорная вероятность модели μ_k. Коэффициенты результирующей регрессии $y \sim X$ оцениваются по выборке из совместного апостериорного распределения вероятностей для коэффициентов регрессий, соответствующих моделям μ_k, $k \in B$. Значимость каждого отдельного геномного маркера $j \in \bigcup_{k \in B} A_{(k)}$ может быть определена как сумма апостериорных вероятностей моделей из множества $M(c)$, в которые входит этот маркер.

Таким образом, в отличие от обычного регрессионного подхода к оценке значимости индивидуальных мутаций, процедура на основе алгоритма MOSS позволяет ранжировать и сравнивать модели в соответствии с их апостериорными вероятностями и использовать байесовское усреднение, чтобы оценить значимость отдельных полиморфизмов по нескольким моделям, которые наилучшим образом согласуются с данными.

Результаты и их обсуждение. По итогам анализа популяционной структуры исследуемые образцы были разделены на несколько групп. С помощью сполиготипирования удалось установить генетические семейства наиболее представленных штаммов *M. tuberculosis*: Beijing (63,6 %), T1 (18,9 %), H3 (5,6 %) и T5 (2,8 %).

Корреляционный анализ результатов тестов на чувствительность к противотуберкулезным препаратам позволил проанализировать перекрестную устойчивость между лекарствами (табл. 1).

Таблица 1. **Парные корреляции между результатами лабораторных тестов на чувствительность к противотуберкулезным препаратам**

Препарат	EMB	INH	RIF	PZA	STM	CYCL	ETH	PARA	AMIK	CAPR	OFLO
EMB	1,00	0,90	0,90	1,00	0,80	0,36	0,34	0,25	0,53	0,59	0,55
INH	0,90	1,00	1,00	0,91	0,89	0,36	0,31	0,23	0,48	0,52	0,53
RIF	0,90	1,00	1,00	0,91	0,89	0,37	0,31	0,22	0,48	0,53	0,53
PZA	1,00	0,91	0,91	1,00	0,72	0,38	0,38	0,18	0,38	0,46	0,49
STM	0,80	0,89	0,89	0,72	1,00	0,33	0,27	0,20	0,42	0,45	0,47
CYCL	0,36	0,36	0,37	0,38	0,33	1,00	0,33	0,32	0,27	0,35	0,23
ETH	0,34	0,31	0,31	0,38	0,27	0,33	1,00	0,06	0,33	0,46	0,14
PARA	0,25	0,23	0,22	0,18	0,20	0,32	0,06	1,00	0,07	0,13	0,23
AMIK	0,53	0,48	0,48	0,38	0,42	0,27	0,33	0,07	1,00	0,90	0,57
CAPR	0,59	0,52	0,53	0,46	0,45	0,35	0,46	0,13	0,90	1,00	0,61
OFLO	0,55	0,53	0,53	0,49	0,47	0,23	0,14	0,23	0,57	0,61	1,00

Интерес представляет исследование дисперсий предсказаний, полученных с помощью линейной смешанной модели, и выделение в них составляющей, обусловленной геномными вариациями (proportion of phenotypic variance explained, PVE). Рассматриваемую величину можно охарактеризовать соотношением $k_{SNP} = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$, где σ_g^2 – изменчивость, объясняемая вариациями в геноме, σ_e^2 – изменчивость, объясняемая факторами внешней среды. Эта информация (табл. 2) дает представление о том, насколько достоверно лекарственная устойчивость может быть предсказана по вариациям в имеющихся геномах.

Таблица 2. **Доля дисперсии фенотипа (чувствительный / устойчивый), объясненная изменениями в геноме**

Препараты первого ряда	PVE, %	Стандартная ошибка, %	Препараты второго ряда	PVE, %	Стандартная ошибка, %
INH (изониазид)	99,997	0,021	CYCL (циклосерин)	75,716	12,386
RIF (рифампицин)	99,997	0,021	CAPR (капреомицин)	73,903	11,048
PZA (пиразинамид)	99,997	0,049	AMIK (амикацин)	69,831	11,925
STM (стрептомицин)	99,997	0,036	OFLO (офлоксацин)	58,682	12,922
EMB (этамбутол)	97,119	1,695	ETH (этионамид)	45,680	24,906
			PARA (кислота парааминосалициловая)	29,998	17,010

Основываясь на приведенных данных можно заключить, что одни лишь геномные полиморфизмы не могут полностью объяснить наблюдаемую лекарственную устойчивость к большинству препаратов второго ряда, что следует учитывать при интерпретации результатов полногеномного анализа ассоциаций.

Для поиска мутаций, связанных с лекарственной устойчивостью, на этапе предобработки данных была выполнена фильтрация геномных вариаций по качеству их детектирования. Также из рассмотрения исключались протяженные вставки/удаления и мутации, присутствующие менее чем в 1 % образцов и повторяющиеся столбцы матрицы генотипов. Приведем обзор качества предсказаний, полученных с помощью рассмотренных методов анализа ассоциаций на предобработанных наборах данных, содержащих 765 полиморфизмов и число штаммов от 34 до 132 (табл. 3). Для оценки обобщающей способности моделей использовался метод перекрестной проверки и вычислялись метрики, общепринятые для оценки качества классификации: точность – доля истинно устойчивых образцов относительно всех объектов, которые классификатор отнес к классу лекарственно-устойчивых; полнота – доля найденных классификатором лекарственно-устойчивых образцов относительно всех истинно устойчивых образцов в выборке; *F*-мера – среднее гармоническое между точностью и полнотой; правильность – доля объектов, по которым классификатор принял правильное решение.

Таблица 3. **Сравнительный анализ качества предсказаний, полученных с помощью многомаркерного анализа ассоциаций**

Препарат	Количество устойчивых образцов	Количество чувствительных образцов	Метод	Точность	Полнота	*F*-мера	Правильность
OFLO	69	63	MOSS	0,929	0,752	0,831	0,840
			LMM	0,557	1	0,715	0,583
			Логистическая регрессия	0,971	0,986	0,978	0,977
EMB	102	30	MOSS	0,962	0,981	0,972	0,955
			LMM	1	0,108	0,195	0,311
			Логистическая регрессия	0,990	1	0,995	0,992
INH	106	26	MOSS	1	0,981	0,990	0,985
			LMM	1	1	1	1
			Логистическая регрессия	1	1	1	1
PZA	28	6	MOSS	1	1	1	1
			LMM	0,966	1	0,983	0,971
			Логистическая регрессия	1	1	1	1
RIF	106	26	MOSS	1	0,869	0,930	0,895
			LMM	1	1	1	1
			Логистическая регрессия	1	1	1	1
STM	110	22	MOSS	1	0,954	0,977	0,962
			LMM	1	0,991	0,995	0,992
			Логистическая регрессия	1	1	1	1
AMIK	59	63	MOSS	1	0,847	0,917	0,926
			LMM	0,484	1	0,652	0,484
			Логистическая регрессия	1	0,932	0,965	0,967

Окончание табл. 3

Препарат	Количество устойчивых образцов	Количество чувствительных образцов	Метод	Точность	Полнота	*F*-мера	Правильность
CAPR	66	51	MOSS	1	0,731	0,845	0,848
			LMM	0,564	1	0,721	0,564
			Логистическая регрессия	1	1	1	1
CYCL	46	70	MOSS	0,502	0,453	0,477	0,604
			LMM	0,397	1	0,568	0,397
			Логистическая регрессия	1	0,978	0,989	0,991

Следует отметить, что результирующие множества значимых мутаций, получаемых разными методами, отличались по количеству и составу признаков. Наименьшее число значимых признаков при достаточно хороших показателях качества классификации было получено алгоритмом MOSS. Наилучшие результаты по итогам перекрестной проверки показала регуляризованная логистическая регрессия. Однако, несмотря на меры по сокращению числа анализируемых с помощью регрессии признаков (фильтрация данных, несколько итераций запуска алгоритма для отбора значимых признаков), количество значимых коэффициентов оставалось неоправданно большим, что может свидетельствовать о включении шумовых признаков в результирующее множество. Тем не менее сопоставление найденных геномных маркеров лекарственной устойчивости с уже используемыми в современных системах молекулярно-генетической диагностики показало, что все применяемые нами методы (логистическая регрессия, линейная смешанная модель, метод на основе ориентированного на моду стохастического поиска), как правило, приписывали найденным маркерам наибольшую значимость.

Заключение. В настоящей работе предложена методология, позволяющая исследовать взаимосвязь мутаций в геномах микроорганизмов с развитием лекарственной устойчивости к противомикробным препаратам. Описанная методология была использована в проекте CRDF OISE-14-60497-1 для исследования 136 штаммов микобактерий туберкулеза, выделенных у пациентов из Беларуси с различными формами туберкулеза легких.

Список использованной литературы

1. *Agresti, A.* An Introduction to Categorical Data Analysis / A. Agresti. – Hoboken, 2002.

2. Principal components analysis corrects for stratification in genome-wide association studies / A. L. Price [et al.] // Nat. Genet. – 2006. – N 38 (8). – P. 904–909.

3. *Zou, H.* Regularization and variable selection via the elastic net / H. Zou, T. Hastie // J. R. Statist. Soc. B. – 2005. – N 67 (2). – P. 301–320.

4. *Zhou, X.* Genome-wide efficient mixed-model analysis for association studies / X. Zhou, M. Stephens // Nature Genetics. – N 44 (7). – 2012. – P. 821–824.

5. Efficient control of population structure in model organism association mapping / H. M. Kang [et al.] // Genetics. – N 178 (3). – 2008. – P. 1709–1723.

6. *Dobra, A*. The mode-oriented stochastic search (MOSS) for log-linear models with conjugate priors / A. Dobra, H. Massam // Statistical methodology. – 2010. – N 7. – P. 240–253.

Поступила в редакцию 21.12.2015