ИДЕНТИФИКАЦИЯ МИКОБАКТЕРИИ ТУБЕРКУЛЕЗА И НЕТУБЕРКУЛЕЗНЫХ МИКОБАКТЕРИЙ НА ОСНОВЕ ГЕНОМНЫХ ДАННЫХ

А. В. Гузаревич

Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь, fpm.guzarevіAV@bsu.by
Научный руководитель — А. В. Тузиков, доктор физико-математических наук, профессор

В статье исследуются методы дифференциации микобактерий туберкулеза и родственных им нетуберкулезных микобактерий на основе последовательностей ДНК. Описываются результаты, полученные при анализе признаков помощью точного теста Фишера, метода опорных векторов, а также сравнении нуклеотидных последовательностей гена MPB64.

Ключевые слова: микобактерия туберкулеза; нетуберкулезные микобактерии; тест Фишера; метод опорных векторов; ген MPB64.

ВВЕДЕНИЕ

Эффективность лечения туберкулеза и заболеваний, вызываемых микобактериями нетуберкулезного комплекса, зависит от быстроты и точности обнаружения соответствующих бактерий в организме человека. Широко применяются традиционные методы идентификации по фенотипическим характеристикам, таким как морфология, скорость роста, пигментация, оптимальный температурный режим и др. [1]. Однако эти методы трудоемки, относительно медленны и могут вызывать задержки и ошибки в терапевтических стратегиях. Применение результатов традиционных лабораторных подходов затрудняется и тем, что организмы, принадлежащие к различным родственным видам, могут обладать одинаковыми фенотипическими признаками. С целью преодоления недостатков известных способов разрабатываются молекулярные методы исследования организмов, в том числе анализ последовательности ДНК.

ВЫБОРКА ДАННЫХ

Выборка данных для исследования включает ДНК-последовательности микобактерий туберкулеза (49 последовательностей) и трех видов нетуберкулезных микобактерий: Mycobacteroides abscessus (45 последовательностей), Mycobacterium avium (34 последовательности), Mycobacterium intracellulare (29 последовательностей). Всего выборка содержит 157 последовательностей ДНК.

ИССЛЕДУЕМЫЕ ПРИЗНАКИ

Признаки ДНК-последовательностей, исследуемые в данной работе, представляют собой k-меры – упорядоченные пары нуклеотидов (A-A, A-C, A-G, A-T, C-A и т.д.), такие, что нуклеотиды, образующие пару, находятся на расстоянии k в геномной последовательности, т.е. между ними расположено k любых нуклеотидов. Исследование было проведено для всех k, не превосходящих 10. Значением признака для ДНК-последовательности считается нормированное количество соответствующих k-меров, встречающихся в данной последовательности. Будем использовать обозначение XkY для признака, где X и Y – символы нуклеотидов (A, C, G или T), k – расстояние между ними в последовательности. Таким образом, в рассмотрении находится $4 \cdot 11 \cdot 4 = 176$ признаков.

ОПИСАНИЕ ИССЛЕДОВАНИЯ И РЕЗУЛЬТАТЫ

Наиболее простыми являются признаки, по которым выборка разделима на числовой прямой, т.е. те, для которых существует такое пороговое значение t данного признака, что для всех геномов туберкулезных микобактерий значение признака меньше (больше) t, а для всех геномов нетуберкулезных микобактерий – больше (меньше) t. Для исследуемой выборки было выявлено два таких признака – C1A (рис. 1) и G1T. По остальным 174 признакам выборки частично или полностью «перекрываются».

Рис. 1. Признак, по которому виды микобактерий разделимы

К каждому признаку был применен двусторонний точный тест Фишера. Результатом работы теста является величина p-value, которая отражает вероятность того, что между значениями признака и видом микобактерии не существует связи. Чем меньше p-value, тем более значимым считается признак. Среди 176 признаков было выявлено 20, для которых p-value не превосходит 0.05, что дает основание считать эти признаки существенными в задаче классификации микобактерий.

Помимо теста Фишера в данной работе к задаче дифференциации микобактерий был применен метод опорных векторов. Классификация была проведена как по каждому признаку по отдельности, так и по всевозможным парам признаков. В экспериментах применялись различные ядра преобразования пространства данных (линейное, полиномиальное, РБФ (радиальная базисная функция), сигмоидное), а также метод опорных векторов без использования преобразования пространства. Наилучшие результаты достигаются с помощью применения ядра РБФ, а именно на 94% пар признаков достигается точность классификации (accuracy) не менее 0.98. На рисунке 2 изображено признаковое пространство для пары k-меров C10A-T6T и расположение объектов выборки в нем.

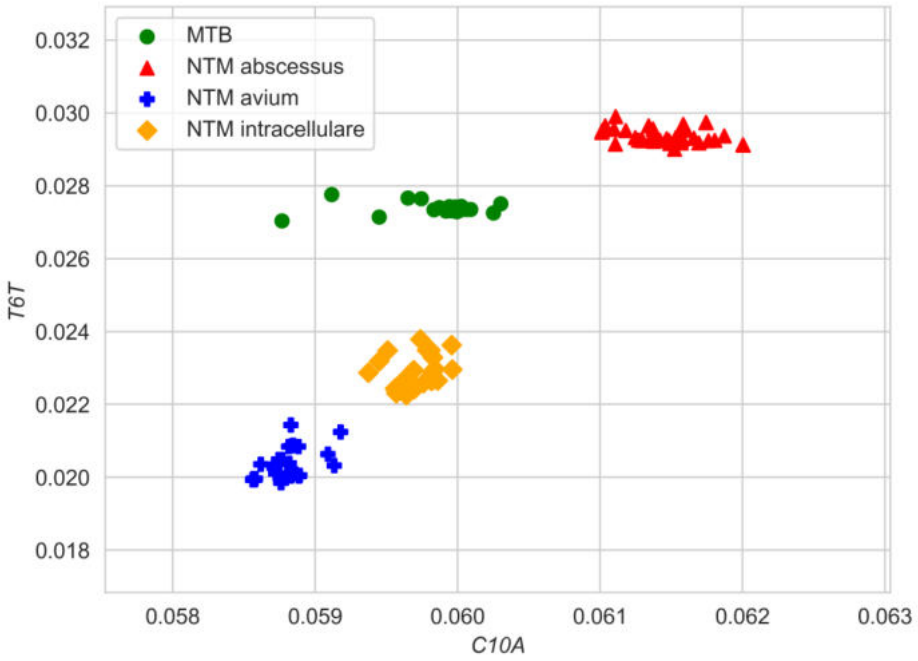

Рис. 2. Пример признакового пространства и расположения объектов выборки в нем

По рисунку 2 видно, что выборки различных видов микобактерий линейно разделимы по данным признакам, что объясняет высокую точность классификации методом опорных векторов.

С помощью описанных выше методов осуществляется поиск признаков в полном геноме бактерий. Однако известны определенные гены, несущие информацию, специфическую для микобактерий туберкулеза [2]. Одним из них является ген MPB64 (другое название – MPT64). Из каждой ДНК-последовательности выборки был «вырезан» участок, соответствующий данному гену. Все последовательности гена MPB64 были ранжированы по их мере сходства с референсной последовательностью этого же

гена микобактерии туберкулеза. В качестве меры сходства двух последовательностей в данной работе использовалась оценка BLAST (Basic Local Alignment Search Tool – инструмент для поиска локальных выравниваний генома) их оптимального выравнивания друг с другом. В результате для каждого из четырех рассматриваемых видов микобактерий был получен некоторый диапазон, в котором лежат значения оценки сходства бактерий этого вида с туберкулезной по гену MPB64 (рис.3).

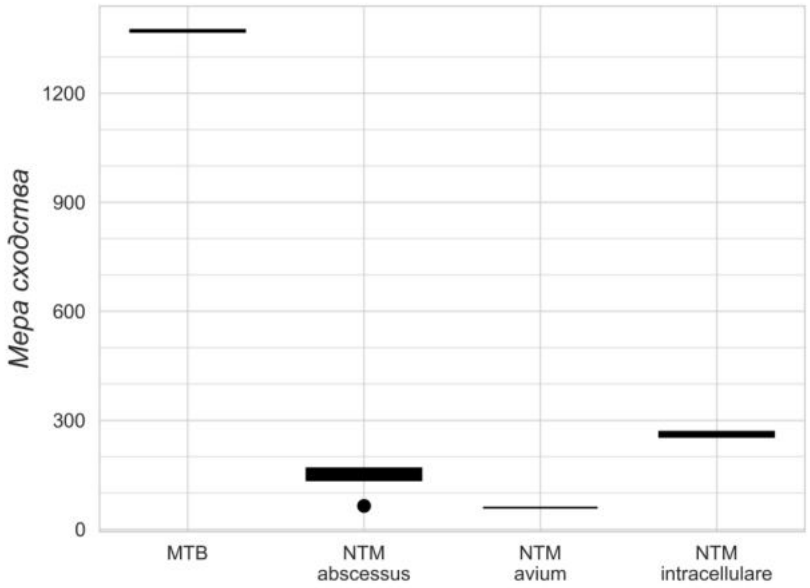

Рис. 3. Диапазоны оценки сходства микобактерий разных видов с туберкулезной по гену MPB64. Кружком обозначен выброс, т.е. значение, существенно отличающееся от всех аналогичных

Как видно по рисунку 3, микобактерии туберкулеза имеют гораздо более высокую оценку выравнивания, чем любая нетуберкулезная микобактерия выборки. Это дает основание считать ген MPB64 значимым в задаче классификации микобактерий. Кроме того, различные виды нетуберкулезных микобактерий демонстрируют разные значения оценки выравнивания, следовательно, можно выдвинуть гипотезу о возможности дифференциации нетуберкулезных микобактерий различных видов между собой по этому же признаку.

Таким образом, в ходе работы были найдены признаки в виде k-меров, которые имеют связь с видом микобактерии, а также исследовано различие последовательностей гена MPB64 между туберкулезными и нетуберкулезными микобактериями. Кроме того, выявлена возможность дифференциации различных видов нетуберкулезных микобактерий на основе k-меров и гена MPB64.

Библиографические ссылки

1. *O'connor J. A. [et al.]* Mycobacterium diagnostics: from the primitive to the promising // British Journal of Biomedical Science. 2015. Vol. 72, № 1. P. 32-41.

2. *Sawatpanich A.[et al.]* Diagnostic performance of the Anyplex MTB/NTM real-time PCR in detection of Mycobacterium tuberculosis complex and nontuberculous mycobacteria from pulmonary and extrapulmonary specimens // CellPress. 2022. Vol.8. e11935.