«Сейфуллин оқулары – 16: Жаңа формациядағы жастар ғылыми – Қазақстанның болашағы» атты халықаралық ғылыми-теориялық конференциясының материалдары = Материалы Международной научно-теоретической конференции «Сейфуллинские чтения – 16: Молодежная наука новой формации – будущее Казахстана». - 2020. - Т.I, Ч.3 - С.149-151

# РАЗРАБОТКА R-СКРИПТА ДЛЯ ОБРАБОТКИ ДАННЫХ МАСС-СПЕКТРОМЕТРИИ

***Голенко Е.С., Исмаилова А.А.***

Данные, полученные с помощью масс-спектрометра и, как правило, всех видов спектрометров, страдают от шума и ошибок, вызванных различными факторами, такими как составная подготовка, искажения и шум, вносимые инструментами анализа. Недавнее применение масс-спектрометрии (МС) к методу нуклеиновых кислот с программируемой белковой матрицей (NAPPA) для идентификации белков неклассическими методами приводит к необходимости более сложного алгоритма распознавания пиков [1]. Техника NAPPA позволяет синтезировать функциональные белки непосредственно из напечатанных ДНК, но сталкивается с трудностями, вызванными наличием мастер-микса и пиков молекул лизата, появляющихся в качестве фона в общих спектрах [2]. Доступен широкий спектр инструментов для анализа белков с обычными масс-спектрами, соответствующими нескольким молекулярным видам. Ни один из них не оптимизирован для вычитания фона. Кроме того, идентификация пиков выполняется путем статистического анализа пиков характеристик, и, таким образом, вычитание фона может изменить результат, удалив характеристические пики [3][4].

Другие недостатки связаны с искажением, расширением пика, насыщением, неправильной калибровкой и различными видами зашумлений. Очистка данных обычно выполняется в два этапа. Наша работа сфокусирована на этапе реализации алгоритмов предварительной обработки, так как именно они представляют наибольший интерес с технической точки зрения. Здесь обсуждается разработка R-скрипта для предварительной обработки данных масс-спектрометрии и ее использования. Предлагаемое программное обеспечение (SCR) будет различать пики белка и фона, позволяя идентифицировать белок.

Были предприняты некоторые попытки анализа имеющихся данных масс-спектрометрии с использованием открытых платформ. На данный момент в открытом доступе находится широкий спектр инструментов для идентификации масс-спектров белка, но исходные образцы должны состоять из нескольких молекулярных видов.

Хотя многие из них предоставляют функции для предварительной обработки, такие как сглаживание, вычитание базовой линии и нормализация, лишь некоторые способны выполнять кластеризацию и анализ данных. К примеру, Dante R написан на R и обеспечивает функцию

кластеризации [5]. Наш скрипт обеспечивает функции предварительной обработки, такие как сглаживание, выделение пиков и нормализация, также может выполнять выравнивание пиков, применять порог и предоставляет функции для вычитания фона, чтобы упростить идентификацию пиков даже при наличии шума. Наконец, его можно использовать в сочетании с простыми алгоритмами интеллектуального анализа данных, такими как k-means, чтобы идентифицировать белки, когда поиск в базе данных в MASCOTT невозможен.

Как входные, так и выходные файлы SCR имеют формат ASCII. Преимущество формата ASCII состоит в том, что он не зависит от платформы, прост в использовании и удобен для чтения, легок и может быть получен независимо от устройства, используемого для сбора данных.

Извлечение пиков определяет наиболее значимое пиковое значение, выбирая реальные пики среди огромного количества шумных пиков. Применяется процедура биннинга, а затем SCR ищет локальный максимум в каждом окне биннинга. Как только эти реальные пики извлечены, SCR может использовать три другие функции:

1. Выравнивание пиков: выравнивание основного пика двух сравниваемых спектров, то есть спектра и шума. SCR может выровнять только основные пики окна с разбивкой. Он ищет максимальный пик в области биннинга и использует его в качестве реферальной точки. Затем все другие основные пики и спектры в одном и том же сечении сдвигаются и выравниваются по этому последнему как в спектрах белка, так и в спектрах шума.

2. Порог: постоянная интенсивность до нуля, если она меньше минимального желаемого значения. Порог не определяется автоматически, и его можно вставить в зависимости от практики и потребностей пользователя.

3. Сведение всех пиков к одному: он используется в сочетании с порогом, а пики равны одному.

Таким образом, результаты могут быть использованы в качестве входных данных для других программных пакетов, которые реализуют алгоритмы интеллектуального анализа данных, мы протестировали его приложение с реализацией R-алгоритма кластеризации k-means. Кроме того, его применение к нетрадиционному обнаружению MS может преодолеть фоновое выражение сигнала, проблемы, с которыми сталкиваются, в частности, при использовании системы NAPPA / SNAP. Для проверки последнего приложения был проведен кластерный анализ на основе различных образцов белка. В частности, использовались три разных набора данных, первый из которых состоял из двух разных спектров белков, полученных с помощью классического метода MC, а второй и третий образцы были составлены из спектров, полученных с помощью модифицированной методики без меток MC на основе SNAP NAPPA. Второй набор данных состоял из известного белка с известным смещением по массиву, в то время как третий набор данных состоял из известного белка, смещенного лишь частично в известных положениях. Последний тест был

выполнен, чтобы оценить различающую мощность SCR в сочетании с реализацией k средних.

В некоторых случаях было трудно различить кластер из одинаковых белков. Чтобы решить эту проблему, выполнялась предварительная обработка вычитания шума, заданная в качестве входных спектров, мастер-микса как спектров шума. Последний может предоставлять наборы данных, которые после кластерного анализа позволяют пользователю различать белковые спектры на основе кластерной вероятности, которые кажутся более дискриминирующими, чем без вычитания шума. Это означает, что распознавание без вмешательства человека и классического анализа МС невозможно даже при использовании предварительной обработки и k-means алгоритмов.

Подводя итог, можно сказать, что, хотя результаты, полученные при первых запусках скрипта, выглядят достоверно, предложное решение требует тестирования с применением обширного диапазона входных данных. Как хорошо известно, применение алгоритмов интеллектуального анализа сильно зависит от проблемы, и возможным выходом из этой проблемы может быть использование SCR в сочетании с другими реализациями алгоритмов кластеризации или классификации, или с более прямым вычитанием фона из сигнала.

## Список литературы


1 Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster, Quantitative mass spectrometry in proteomics: a critical review // Anal Bioanal Chem. – 2007. – №398. –С. 1017-1031.

2 Spera R, LaBaer J, Nicolini C, Mass Spectrometry Detection of Nucleic Acid Programmable Protein Array // Journal of Mass Spectrometry. – 2011. - №46. – С. 960-965.

3 Nicolini C, Bragazzi N, Pechkova E, Nanoproteomics enabling personalized nanomedicine // Adv Drug Deliv Rev. – 2012. –№64. – С.1522-1531.

4 Nicolini C, Labaer J, Functional Proteomics and Nanotechnology-based Microarrays // Pan Stanford Series on Nanobiotechnology. – 2010. – №2. –С. 1-308.

5 Polpitiya AD, Qian WJ, Jaitly N, Petyuk VA, Adkins JN, et al., DAnTE: a statistical tool for quantitative analysis of omics data // Bioinformatics. – 2008. – №24. – С. 1556-1558.