Правительство Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный университет»
(СПбГУ)

УДК 57:51-76 57.02:001.57
Рег № НИОКТР
АААА-А20-120013090048-8
Инв.№ 75307362

УТВЕРЖДАЮ
Начальник Управления
научных исследований СПбГУ
_____ Е.В. Лебедева
« » _____20___ г.

ОТЧЁТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

Разработка вычислительных методов для идентификации вторичных метаболитов растений с использованием данных масс-спектрометрии

(заключительный)

по гранту РФФИ
№ 20-04-01096 от 19.02.2020

Руководитель НИР,
старший научный сотрудник,
кандидат физико-математических наук А.А. Гуревич

Санкт-Петербург
20___

РЕФЕРАТ

Отчет 12 с., 5 рис., 2 источн.

В рамках данного проекта получило свое развитие перспективное направление анализа данных тандемной масс-спектрометрии вторичных метаболитов растений. Разработаны вычислительные подходы для высокопроизводительной идентификации вторичных метаболитов по масс-спектрам и базе химических соединений. Апробация этих подходов на реальных данных показала их эффективность и высокую точность при идентификации различных классов вторичных метаболитов (нерибосомные пептиды, поликетиды, сахариды, сапонины и др.) как растительного, так и бактериального, и грибного происхождения.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Вторичные метаболиты растений лежат в основе многих лекарственных препаратов включая антибиотики, иммунодепрессанты, витамины и другие. Современные технологии масс-спектрометрии позволяют за короткое время произвести сканирование образцов с тысячами метаболитов, получая в результате огромные объемы данных высокого разрешения. Несмотря на то, что эти данные представляют собой перспективный источник для открытия новых соединений, их интерпретация остается узким местом, требующим развития соответствующих вычислительных методов.

Существенные шаги в разработке подобных биоинформатических программ были предприняты в последние годы, однако они все еще не могут быть эффективно применены к обработке масс-спектров вторичных метаболитов растений, так как либо ориентированы только на строго определенные классы химических соединений, либо работают слишком медленно или только с небольшими базами данных. Кроме того, большинство из существующих методов выполняют только так называемую “стандартную” идентификацию масс-спектров, то есть выявление тех из них, которые относятся к известным молекулам из заданной базы химических соединений. В то время как стандартная идентификация крайне важна для предотвращения “переоткрывания” известных молекул, что значительно экономит ресурсы и время ученых, она принципиально не позволяет найти новые соединения в данных, что является конечной целью большинства подобных исследований.

Данный проект направлен на разработку новых подходов к высокопроизводительному анализу данных тандемной масс-спектрометрии вторичных метаболитов растений. В рамках проекта получили развитие методы идентификации метаболитов по данным масс-спектрометрии и базам известных химическим соединений. Разработан метод “вариативной” идентификации, позволяющей идентифицировать спектры не только известных соединений, но и их вариантов, не представленных в базе данных.

Вычислительные методы были созданы в тесном сотрудничестве опытных разработчиков научного программного обеспечения и их конечных пользователей – исследователей вторичных метаболитов. Оба подхода доступны в виде приложения для командной строки и удобного веб сервиса на GNPS, крупнейшей международной платформе по анализу масс-спектров вторичных метаболитов (Wang et al, Nature Biotechnology, 2016). Мы рассчитываем, что созданные методы будут активно использоваться другими научным группами, изучающими вторичный метаболизм, и помогут быстрее находить потенциально полезные биоактивные вещества природного происхождения.

ОСНОВНАЯ ЧАСТЬ ОТЧЕТА НИР

Разработка метода идентификации метаболитов по данным масс-спектрометрии и базам известных химическим соединений

MolDiscovery создан на базе разработанного нами ранее инструмента Dereplicator+ (Mohimani et al, 2018), но использует новые алгоритмические и программные решения, которые существенно улучшили производительность и точность программы. Ключевые использованные подходы подробно описаны ниже, а также в статье о molDiscovery (Cao et al, 2021). Программа molDiscovery реализована в виде свободно доступных для академических исследователей приложения командной строки и веб-сервиса на международной платформе GNPS.

При разработке molDiscovery был предложен ряд новых методов и подходов. Для построения графа фрагментации молекулы в molDiscovery применен новый подход, основанный на алгоритме Тарьяна и Хопкрофта (Hopcroft & Tarjan, 1973) для разбиения графов на компоненты связности. Этот подход позволил сделать построение графа фрагментации значительно более эффективным как по времени работы, так и по использованию оперативной памяти. Так, сравнение с Dereplicator+ показало ускорение почти на два порядка и уменьшение потребления оперативной памяти более чем в пять раз для молекул массой более 600 Да.

Для поддержки баз данных большого размера было введено индексирование предобработанной базы данных. Индексирование позволяет molDiscovery быстро загружать в оперативную память только необходимые записи, в то время как в Dereplicator+ в память загружается вся предобработанная база данных. Сравнение работы molDiscovery и Dereplicator+ при использовании базы данных из 82 тысяч соединений показало уменьшение времени работы в семь раз и уменьшение потребляемой оперативной памяти более чем в 100 раз.

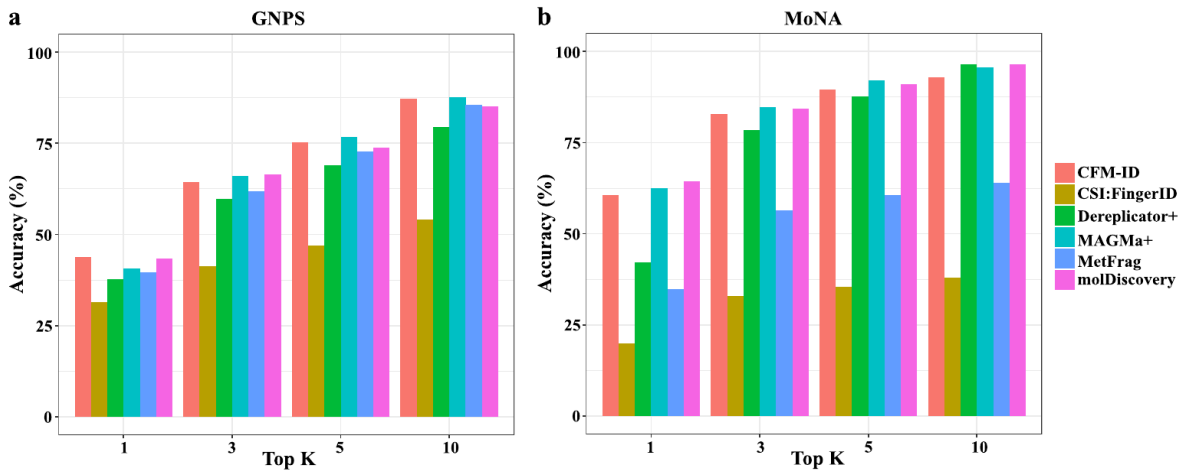

Рис. 1. Точность различных инструментов среди первых K=1, 3, 5, и 10 идентификаций (a) при поиске 194 спектров из спектральной библиотеки GNPS в базе данных DNP (77,057 молекул) и (b) при поиске 342 спектров из библиотеки MoNA в базе данных из 10,124 молекул из MoNA (см. Cao et al., 2021).

В molDiscovery внедрен новый метод идентификации, основанный на вероятностной модели. Прежний метод, разработанный для продукта Dereplicator+, имел ряд серьезных недостатков: он не учитывал, что при фрагментации молекулы в масс-спектрометре разные молекулярные связи (например, соединяющие два атома углерода или атом углерода с атомом кислорода) имеют разную вероятность разрыва; что разрыв молекулярной связи при фрагментации линейного участка молекулы более вероятен, чем разрыв пары связей, необходимый для фрагментации циклического участка; что более интенсивные пики в экспериментальном спектре должны соответствовать более вероятным шаблонам фрагментации. Кроме того, прежний метод не был устойчив относительно размера молекулы и числа пиков в экспериментальном спектре, что приводило к смещению результатов в сторону молекул большого размера и спектров с большим числом пиков. Новый метод основан на вероятностной модели, и учитывает все перечисленные недостатки. Параметры модели были получены при помощи методов машинного обучения на обучающей выборке аннотированных спектров из спектральной библиотеки GNPS (Wang et al, 2016). Применение нового метода позволило увеличить точность работы molDiscovery (см. Рис. 1).

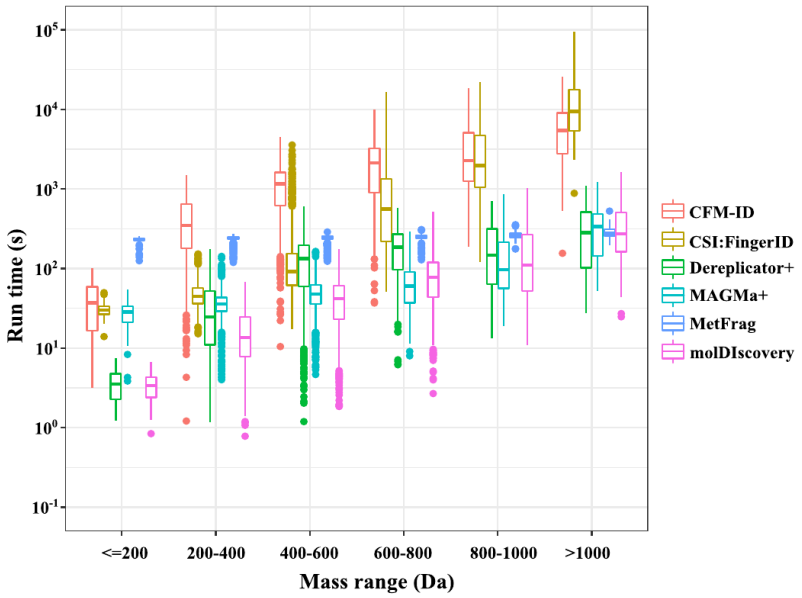

Рис. 2. Время работы (вертикальная ось) разных методов для разных диапазонов массы вещества (горизонтальная ось) без учета времени предобработки базы данных. Центральная линия ящичковой диаграммы соответствует медиане. Границы ящиков соответствуют первому и третьему квартилю. Длина уса составляет полтора межквартильных расстояния.

В рамках проекта был проведен ряд вычислительных экспериментов на аннотированных масс-спектрометрических данных, в которых работа molDiscovery сравнивалась как с Dereplicator+, так и ведущими сторонними программами для идентификации вторичных метаболитов CFM-ID (Allen, et al, 2015), CSI:FingerID (Dührkop, et al, 2015), MetFrag (Wolf,et al., 2010) и MAGMa+ (Verdegem, et al, 2016). Сравнения показали значительное превосходство нового подхода по времени работы и

потребляемой памяти над всеми конкурирующими методами (Рис. 2). При этом процент правильных идентификаций у molDiscovery оказался существенно выше, чем у Dereplicator+ и сравним или лучше, чем у сторонних методов (Рис. 1). Таким образом, molDiscovery позволяет выполнять обработку крайне больших наборов данных (миллионы масс-спектров против сотен тысяч химических соединений) за относительно небольшое время (дни/недели) и выдавать результат сравнимый или превосходящий по точности лидирующие в области подходы, требующие для обработки такого объема данных месяцы и годы. Насколько нам известно, molDiscovery стал первым и единственным на данный момент высокопроизводительным методом для идентификации широкого круга вторичных метаболитов в данных масс-спектрометрии по базам химических структур.

Разработка метода вариативной идентификации метаболитов

На базе molDiscovery и VarQuest был создан новый инструмент для вариативной идентификации метаболитов посредством поиска масс-спектров по базе данных, varDiscovery (изначально запланированное в заявке название было PlantVarQuest; разработанный в рамках первого года проекта прототип имел название VarQuest+). Этот инструмент включил в себя как новый метод идентификации, так и основные ускорения базовой программы, внедренные в рамках создания molDiscovery (см. выше). Кроме того, в него добавлена функциональность вариативной идентификации по аналогии с VarQuest (Gurevich et al, 2018), но расширенная с пептидных метаболитов на метаболиты произвольной химической структуры. Вариативная идентификация позволяет поставить в соответствие спектру молекулу из базы данных, отличающуюся от истинной некоторой модификацией (молекулу-вариант). В предположении, что модификация является точечной, соответствующая ей разница масс применяется к одной из вершин молекулярного графа. Затем молекулярный граф сравнивается со спектром при помощи метода идентификации разработанного в рамках molDiscovery.

В случае вариативной идентификации пространство поиска увеличивается на несколько порядков по сравнению со стандартной идентификацией. Можно выделить две основные причины: во-первых, при вариативной идентификации для каждого спектра многократно увеличивается число молекул-кандидатов. Это происходит из-за того, что допустимая разница в массе спектра и молекулы-кандидата должна быть не меньше массы предполагаемой модификации, в то время как в стандартной идентификации она не превышает погрешности прибора. Во-вторых, зная массу модификации, нужно перебрать все возможные позиции в молекуле-кандидате для получения молекул-вариантов, которые затем и будут сравниваться с экспериментальным спектром. Все вместе, это приводит к значительному увеличению времени работы программы в вариативном режиме.

Для повышения точности вариативной идентификации и уменьшения времени работы программы, при создании varDiscovery был разработан и внедрен ряд новых методов. Так, в varDiscovery разработан новый метод фильтрации молекул-кандидатов. Как было показано в (Mohimani et al, 2018), для повышения точности идентификации требуется учитывать несколько уровней фрагментации молекулы. Однако, можно заметить, что при этом часть пиков экспериментального спектра как правило объяснена

фрагментами первого уровня. Это связано прежде всего с тем, что при фрагментации молекулы в масс-спектрометре вероятность образования фрагмента первого уровня выше, чем у последующих. В рамках процедуры фильтрации, для дальнейшего рассмотрения отбираются только те молекулы-кандидаты, для которых количество экспериментальных пиков, объясненных фрагментами первого уровня, больше некоторого наперед заданного значения.

В рамках фильтрации реализовано несколько режимов работы, учитывающих смещение пиков экспериментального спектра. Поскольку молекула-кандидат отличается от истинной молекулы за счет некоторой модификации, часть пиков экспериментального спектра будет смещена относительно масс соответствующих им фрагментов на массу этой модификации. В одном из режимов работы рассматриваются только несмещенные пики, в другом только смещенные пики, и в третьем как смещенные, так и не смещенные пики. Позиция модификации при этом не учитывается.

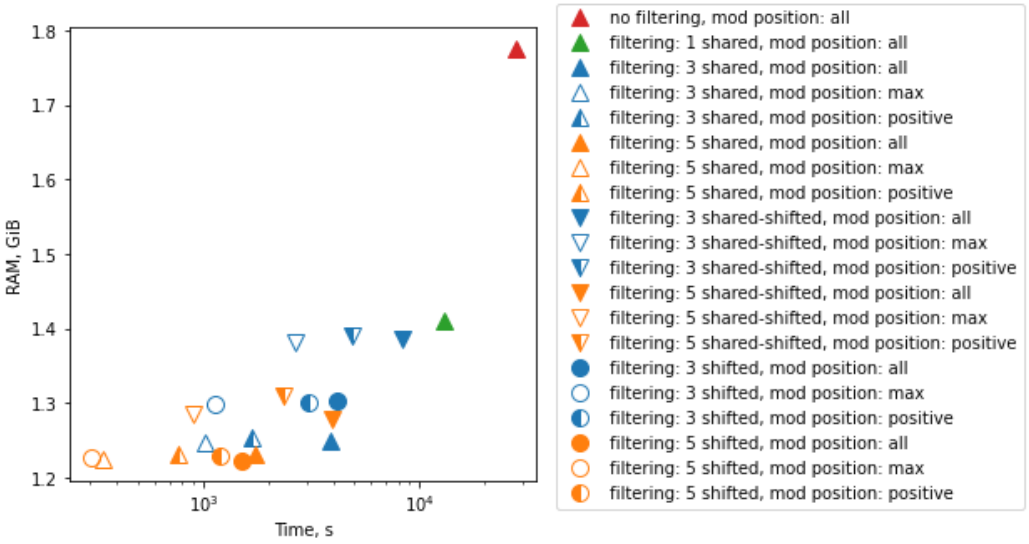

Рис. 3. Время работы varDiscovery (горизонтальная ось) и пиковое потребление памяти (вертикальная ось) для разных режимов фильтрации кандидатов и выбора позиции модификации. Форма точки обозначает режим фильтрации кандидатов: только по несмещенным пикам (прямой треугольник), по смещенным и несмещенным пикам (перевернутый треугольник), только по смещенным пикам (круг). Цвет обозначает минимальное число объясненных пиков для молекулы-кандидата: по крайней мере один пик (зеленый), по крайней мере три пика (синий), по крайней мере пять пиков (оранжевый). Форма заливки обозначает режим выбора позиции для модификации: все позиции (сплошная заливка), позиции для которых число объясненных смещенных пиков превышает число объясненных несмещенных пиков (половинчатая заливка), позиции, для которых модификация объясняет максимальное число смещенных пиков и при этом минимальное число несмещенных пиков (нет заливки). Красный треугольник обозначает режим, в котором фильтрация кандидатов не применялась, и рассматривались все позиции для модификации.

Задача сопоставления фрагментов первого уровня с масс-спектрами была ранее эффективно решена для пептидных молекул в рамках инструмента VarQuest. Ее адаптация к молекулам произвольных химических классов позволила как сократить время работы varDiscovery (Рис. 3), так и повысить точность идентификации за счет отсеивания “плохих” кандидатов (Рис. 4).

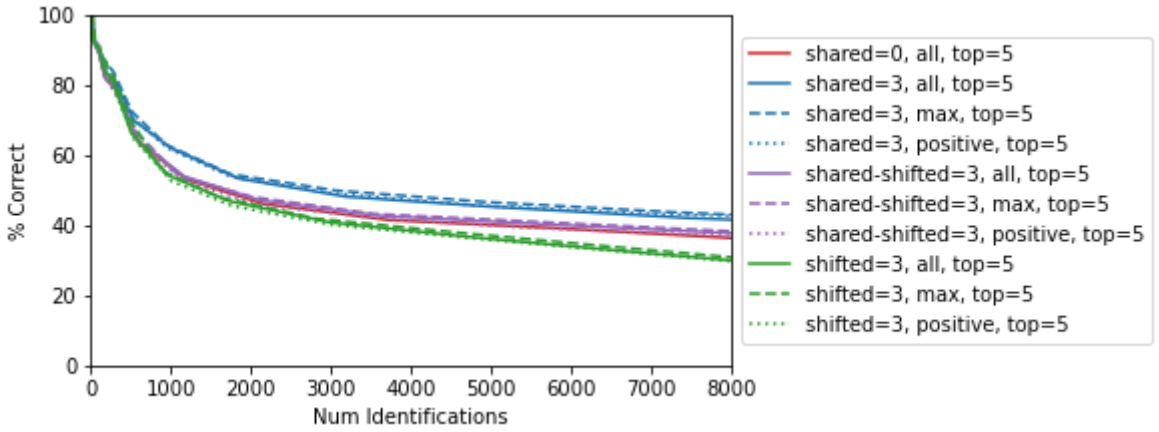

Рис. 4. Процент корректных идентификаций varDiscovery в зависимости от числа идентификаций (отсортированных по скору в порядке невозрастания) для разных режимов фильтрации кандидатов и выбора позиции модификации на датасете из 11,842 аннотированных спектров (Huber et al., 2021). Идентификация считалась корректной, если коэффициент Танимото схожести цифровых отпечатков (ECFP4 fingerprint) молекул составлял не менее 0.5 для хотя бы одного из первых пяти максимальных по скору кандидатов. Цвет кривых обозначает режим фильтрации кандидатов: только по несмещенным пикам (синий), по смещенным и несмещенным пикам (фиолетовый), только по смещенным пикам (зеленый). Штриховка обозначает режим выбора позиции для модификации: все позиции (сплошная), позиции для которых число объясненных смещенных пиков превышает число объясненных несмещенных пиков (пунктирная), позиции, для которых модификация объясняет максимальное число смещенных пиков и при этом минимальное число несмещенных пиков (точечная). Красная кривая соответствует режиму, в котором фильтрация кандидатов не применялась, и рассматривались все позиции для модификации.

Метод построения молекул-вариантов был адаптирован для использования предобработанной базы данных молекул. Новый метод позволяет одновременно с выбором позиции для применения модификации быстро изменять массы фрагментов в графе фрагментации молекулы, хранящемся в предобработанной базе данных.

Кроме того, в рамках метода построения молекул-вариантов была разработана процедура выбора позиции для применения модификации. В рамках этой процедуры анализируется количество смещенных и не смещенных пиков, объясненных фрагментами первого уровня при условии применения модификации к той или иной вершине. Для итогового сравнения выбираются варианты, при которых модификация объясняет максимальное число смещенных пиков и при этом минимальное число несмещенных. Апробация этого метода на аннотированных данных показывает, что его внедрение

приводит к значительному сокращению времени работы инструмента и потребляемой оперативной памяти (Рис. 3) и при минимальном влиянии на точность его работы (Рис. 4).

Реализован инструмент детального вывода сравнения пары молекула-спектр, показывающий скор варианта, полученного применением модификации к каждой из возможных позиций. Значения этих скоров позволяют судить о конкретной позиции, в которой молекула-кандидат отличается от молекулы, соответствующей спектру (Рис. 5).

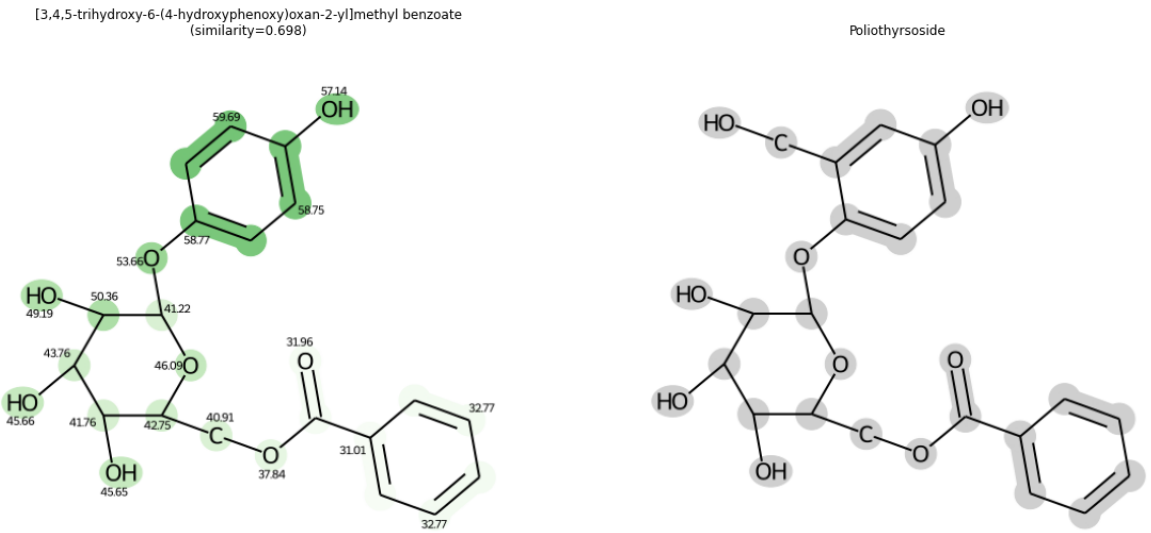

Рис. 5. Результат детализированного вывода varDiscovery. На левой панели представлена молекула-вариант ([3,4,5-trihydroxy-6-(4-hydroxyphenoxy)oxan-2-yl]methyl benzoate), на правой панели – истинный ответ (Poliothyrsoside). Масса предполагаемой модификации составляет 30.01 Да. Вершины метаболитного графа, соответствующего молекуле, выделены цветом. На левой панели каждая вершина аннотирована скором, интенсивность цвета вершины пропорциональна скору при условии применения модификации к данной вершине. Видно, что varDiscovery корректно предсказывает позицию предполагаемой модификации молекулы-варианта.

Вариативная идентификация является гораздо менее проработанной областью. Наиболее близкими аналогами varDiscovery являются VarQuest (Gurevich et al, 2018) и аннотирование вариантов через Молекулярные Сети (Wang et al, 2016). Однако у обоих аналогов есть серьезные ограничения: область применимости VarQuest ограничивается лишь пептидными метаболитами. Молекулярные сети аннотируют спектры сверяя их с известными спектрами веществ, что требует присутствия варианта вещества в спектральной библиотеке. Это условие оказывается не всегда выполнено ввиду ограниченности размера спектральных библиотек. VarDiscovery же работает со всеми классами вторичных метаболитов и не требует наличия спектра известного соединения. Кроме того, в отличие от молекулярных сетей, инструмент varDiscovery способен также предсказывать позицию модификации молекулы-варианта.

ЗАКЛЮЧЕНИЕ

В рамках гранта получило свое развитие перспективное направление анализа данных тандемной масс-спектрометрии вторичных метаболитов:

1. Разработан подход к высокопроизводительной идентификации вторичных метаболитов по масс-спектрам и базе химических соединений. Подход опубликован в журнале Nature Communications (Cao et al, 2021) и реализован в программе molDiscovery, доступной в виде приложения командной строки и веб-интерфейса на международной платформе GNPS (https://gnps.ucsd.edu/). Мы показали эффективность и высокую точность molDiscovery при идентификации различных классов вторичных метаболитов (нерибосомные пептиды, поликетиды, сахариды, сапонины и др.) как растительного, так и бактериального, и грибного происхождения (см. Cao et al, 2021).
2. Продемонстрирована практическая польза от применения molDiscovery в реальных метаболомных исследованиях. MolDiscovery идентифицировал в шесть раз больше уникальных соединений по сравнению с существующими аналогами при анализе более 8 миллионов масс-спектров из общедоступных данных с платформы GNPS против базы с химическими структурами более 700 тысяч вторичных метаболитов (см. Cao et al, 2021). Применение molDiscovery к набору микробных масс-спектров с известными геномами организмов-продуцентов позволило связать масс-спектры 19 молекул с известными биосинтетическими генными кластерами, отвечающими за их синтез, а также обнаружить 3 ранее неописанных генных кластера (см. Cao et al, 2021).
3. На базе molDiscovery разработан подход к вариативной идентификации вторичных метаболитов по масс-спектрам и базе химических соединений. В отличии от существующих аналогов, подход позволяет выполнять высокопроизводительный анализ (миллионы масс-спектров и миллионы соединений) и работает со всеми классами вторичных метаболитов. Подход реализован в программе varDiscovery, подготовлены статья и заявка на регистрацию программы для ЭВМ.

ПУБЛИКАЦИИ

1. MolDiscovery: learning mass spectrometry fragmentation of small molecules; Cao Liu (Carnegie Mellon University), Guler Mustafa (Carnegie Mellon University), Tagirdzhanov Azat M. (Saint Petersburg State University, Saint Petersburg Electrotechnical University “LETI”), Lee Yi-Yuan (Carnegie Mellon University), Gurevich Alexey (Saint Petersburg State University), Mohimani Hosein (Carnegie Mellon University); Nature Communications, T. 12. C. 3718; 17 июня 2021; https://doi.org/10.1038/s41467-021-23986-0
2. VarQuest+: modification-tolerant database search of secondary metabolites mass spectra; Tagirdzhanov Azat M. (Saint Petersburg State University, Saint Petersburg Electrotechnical University “LETI”), Shcherbin Egor (National Research University Higher School of Economics), Mohimani Hosein (Carnegie Mellon University), Gurevich Alexey (Saint Petersburg State University); BMC Bioinformatics, T. 21, C. 567. https://doi.org/10.1186/s12859-020-03838-2; Статья в материалах конференции “Bioinformatics: from Algorithms to Applications (BiATA)”, 27–28 июля 2020