Санкт-Петербургский государственный университет

ЯРЦЕВ Дмитрий Александрович

Выпускная квалификационная работа

Разработка метода деконволюции тандемных масс-спектров с фрагментацией методом ультрафиолетовой диссоциации

Уровень образования: бакалавриат
Направление *01.03.01 «Математика»*
Основная образовательная программа *СВ.5000.2017 «Математика»*

Научный руководитель:
Вяткина Кира Вадимовна,
к.ф-м.н., зав. кафедрой
биоинформатики и
математической биологии
СПбАУ РАН им. Ж.И.
Алфёрова

Рецензент: Иванов Марк
Витальевич, к.ф-м.н.,
научный сотрудник
лаборатории физико-
химических методов
исследования Института
химической физики им. Н.Н.
Семенова Российской
академии наук

Санкт-Петербург
2021

Введение

Масс-спектрометрия белков — это метод идентификации и анализа молекул при помощи прибора, называемого масс-спектрометром. На практике анализ масс-спектров белков и пептидов представляет собой сложную задачу, поскольку сложность технологии и ошибки, возникающие в ходе эксперимента, приводят к низкой воспроизводимости результатов и появлению различий в масс-спектрах, снятых с одного и того же образца.

В масс-спектрометрии существуют два основных подхода — bottom-up и top-down [1, 2]. При подходе bottom-up белок предварительно расщепляется на короткие пептиды, в то время как в соответствии с технологией top-down белковая молекула анализируется целиком. Непосредственным результатом работы масс-спектрометра являются так называемые "сырые" масс-спектры, в которых представлена зависимость величины ионного тока от отношения массы к заряду. Путём выделения пиков из них могут получены центрированные масс-спектры, в которых указанная зависимость представлена дискретными пиками. Далее на этапе деконволюции происходит переход от отношения массы к заряду к нейтральным моноизотопным массам.

С использованием различных приборов могут быть получены масс-спектры низкого, высокого и сверхвысокого разрешения. Для задач количественного анализа, а также идентификации пептидов по базам данных, как правило, достаточно низкого разрешения, однако для решения ряда других задач (в частности, *de novo* секвенирования белков) необходимо использовать высокое или сверхвысокое разрешение.

Технологический прорыв в 2000-х годах [3] привёл к возможности получать данные высокого и сверхвысокого разрешения при сравнительно небольших затратах. Такие масс-спектры весьма информативны, но для их анализа необходимы новые эффективные алгоритмы.

При обработке top-down масс-спектром деконволюция является необходимым шагом, в то время как при обработке bottom-up масс-спектров ее обычно не применяют. Однако было показано, что деконволюция bottom-up масс-спектров, снятых с высоким разрешением, делает возможным применение для их обработки алгоритм Twister (описание изложено в [6, 7, 8, 9]), изначально предназначенный для de novo секвенирования белков. В то же время алгоритмы деконволюции, предназначенные для обработки top-down масс-спектров, не учитывают особенностей bottom-up данных. Тем самым обусловлена необходимость их адаптации к этому случаю.

В данной работе мы предлагаем усовершенствованную версию алгоритма MS-Deconv, предложенного в [4]. Этот алгоритм изначально был предназначен для деконволюции top-down масс-спектров высокого разрешения. Наша цель заключалась в том, чтобы адаптировать алгоритм для работы с bottom-up масс-спектрами высокого и сверхвысокого разрешения. Для этого необходимо было учесть особенности, характерные для масс-спектров пептидов и так называемой "тонкой структуры" изотопных кластеров (fine isotopic structure). Мы внесли в алгоритм изменения, улучшающие результаты обработки масс-спектров пептидов и обеспечивающие возможность его применения к данным сверхвысокого разрешения.

Постановка задачи

Центрированный тандемный белковый масс-спектр представляет из себя набор пиков. Каждый пик характеризуется двумя координатами: m/z — отношением массы к заряду — и I — своей интенсивностью. Задача алгоритма деконволюции — перейти от отношения массы к заряду к нейтральным массам. Это делается за счёт выявления изотопных кластеров. Изотопный кластер — это набор пиков, порождённых ионами с одинаковым атомным составом и зарядом, но разным набором изотопов. В масс-спектрах высокого разрешения первые координаты пиков кластера образуют арифметическую прогрессию с разностью примерно $\frac{m_n}{z}$, где z — заряд иона, породившего кластер, m_n — масса нейтрона. Каждый отступ на один шаг прогрессии соответствует одному дополнительному нейтрону среди изотопов. В реальности добавление одного нейтрона оказывает немного разный эффект на массу различных элементов. Это наблюдение отражает тонкая изотопная структура. Она видна только на масс-спектрах сверхвысокого разрешения, где пики не образуют арифметическую прогрессию, но сконцентрированы возле её членов.

Наш алгоритм планируется применять для bottom-up масс-спектров высокого и сверхвысокого разрешения. Это позволяет сделать дополнительные предположения об устройстве масс-спектра. Во-первых, мы можем считать, что заряд иона принимает только небольшие значения ($z = 1$, 2 или 3). Во-вторых, можно предполагать, что базовый пик кластера (т.е. пик, имеющий наибольшую интенсивность) соответствует моноизотопной массе — массе в случае, когда ион не содержит тяжёлых изотопов. В то же время всякий алгоритм, работающий со масс-спектрами сверхвысокого разрешения, должен учитывать тонкую изотопную структуру, поэтому мы не можем сделать предположение, что каждый член арифметической прогрессии в экспериментальном кластере представлен лишь одним пиком.

Отметим, что задача обнаружения кластеров имеет простое переборное решение, в то время как задача отбора кластеров долгое время оставалась плохо изученной. Предлагаемый в [4] подход — это один из способов получить точное решение задачи отбора. Это положительно выделяет MS-Deconv среди других алгоритмов деконволюции, где жадные алгоритмы отбора позволяют найти лишь приближённое решение задачи.

Описание алгоритма

Следуя идеям MS-Deconv, наш алгоритм состоит из двух частей: обнаружения ионов-кандидатов и их отбора. На шаге обнаружения перебираются все ионы с точностью до атомного состава, для которых базовый пик изотопного кластера близок по величине m/z к одному пиков экспериментального масс-спектра, имеющих интенсивность не ниже уровня шума. В качестве уровня шума принимается значение, равное 10% от интенсивности самого высокого пика исследуемого масс-спектра. Для каждого из проанализированных изотопных кластеров вычисляется значение оценочной функции (о.ф.). Чем выше оценочная функция кластера, тем лучше он соотносится с масс-спектром. Для каждого рассмотренного экспериментального пика во множестве кандидатов остаётся лишь пять кластеров, имеющих такой базовый пик. При этом кластер должен удовлетворять набору условий, которые будут приведены ниже, а в случае их соблюдения предпочтение отдаётся кластерам с наибольшим значением оценочной функции. После фильтрации остаётся лишь решить задачу выбора нескольких кластеров с наибольшим суммарным значением о.ф. Эта задача сводится к поиску пути наибольшего веса в ориентированном ациклическом графе.

Обнаружение кандидатов

На шаге обнаружения кандидатов мы рассматриваем ионы, которые потенциально могли породить изотопные кластеры в экспериментальном масс-спектре, и вычисляем их оценочные функции. Это делается при помощи полного перебора всех последовательностей аминокислот с моноизотопной массой до Mz_0, где M — наибольшая из масс экспериментальных кандидатов на роль базового пика, а z_0 — наибольшее допустимое значение заряда. Благодаря тому, что аминокислотные фрагменты с одинаковым атомным составом имеют один и тот же теоретический изотопный кластер, достаточно перебирать их с точностью до разложения на атомы. Эта идея позволяет провести перебор за время $O((Mz_0)^5)$, так как в состав аминокислот входят только пять химических элементов: водород, углерод, кислород, азот и сера. Отметим, что в экспериментах цистеин подвергается модификации, называемой карбамидометилированием, изменяющей его массу на 57.021 Да. Наш перебор предполагает, что все цистеины модифицированы. Помимо этого, из перебора исключены аминокислоты, имеющие тот же атомный состав, что и димеры из других аминокислот (например, аргинин N имеет тот же атомный состав, что и димер GG), и аминокислоты, имеющие гомологичные (например, треонин T, являющийся гомологом серина S). Вместо последних отдельно перебирается число аминокислот, на месте которых в ионе находятся их гомологи.

После того, как ион был найден, необходимо сгенерировать его теоретический изотопный кластер. В теоретическом изотопном кластере по горизонтали откладывается отношение массы изотопа к его заряду, а по вертикали — вероятность появления изотопа с такой массой. Для генерации теоретического масс-спектра мы используем алгоритм EMass, принимающий на вход атомный состав аминокислотного фрагмента. Мы приведём краткое описание этого алгоритма, а более подробное изложено в [5]. Будем считать, что каждый атом принимает ту или иную изотопную форму независимо от других атомов иона. Теперь рассмотрим для каждого из пяти атомов его производящую функцию. Производящая функция атома — это полиномиальная функция вида $f(x) = \sum p_i x^{m_i}$, где p_i — это частота появления изотопа в природе, а m_i — его масса. В отличие от оригинального MS-Deconv, наши производящие функции учитывают тонкую изотопную структуру, поэтому разница $m_i - m_{i-1}$ слегка различается у разных элементов. Производящая функция аминокислоты и полимера определяется аналогичным образом. Имея набор пиков, легко построить его производящую функцию, и наоборот. Поскольку изотопы разных атомов встречаются независимо, при объединении двух групп атомов в одну их производящие функции перемножаются. При помощи бинарного возведения в степень производящую функцию мономера X_n элемента X можно вычислить за время $O(n^2)$ (или $O(n \log n)$, если использовать быстрое преобразование Фурье). Получившиеся пять

производящих функций затем нужно перемножить между собой. При вычислениях без округления это привело бы к значительному росту числа ненулевых коэффициентов. Чтобы этого избежать, применяются два вида округления. Во-первых, из функции удаляются все мономы $p_i x^{m_i}$, у которых значение p_i ниже определённого порога. Во-вторых, все значения m_i хранятся лишь с определённой точностью. Благодаря этим приёмам число пиков в теоретическом масс-спектре остаётся небольшим даже при учёте тонкой изотопной структуры, что позволяет не задумываться об асимптотике.

Теоретический изотопный кластер дополнительно смещается вправо в зависимости от того, какую природу имеет ион. Для b-ионов ко всем значениям m/z добавляется масса одного протона. Для y-ионов увеличение то же самое, за исключением того, что при генерации теоретического масс-спектра к атомному составу иона добавляются два атома водорода и один атом кислорода (представляющие атомный состав молекулы воды).

Последним шагом перед вычислением оценочной функции является сопоставление теоретическому изотопному кластеру экспериментальных пиков. Для этого каждому из теоретических пиков приписывается ближайший к нему по координате m/z экспериментальный. Заметим, что при работе с масс-спектрами высокого, а не сверхвысокого разрешения, один и тот же экспериментальный пик может быть сопоставлен сразу нескольким теоретическим. В таком случае его интенсивность делится между теоретическими пиками пропорционально их частотности. Если же, наоборот, вблизи теоретического пика нет ни одного экспериментального, в этой точке создаётся фиктивный экспериментальный пик интенсивности 0. Получившиеся два списка пиков затем передаются на вход оценочной функции.

Оценочная функция

Для оценки качества полученных результатов в алгоритме используется специальная функция. Оценочная функция принимает на вход пару $(E, \tilde{E})$ из экспериментального кластера E и теоретического кластера $\tilde{E}$. Прежде чем начать их сравнение, необходимо отмасштабировать теоретический кластер, увеличив интенсивности всех пиков в одно и то же число раз. Выбор масштабирующего коэффициента мы обсудим позднее.

Предположим, что теоретический кластер уже отмасштабирован нужным образом. Тогда для каждого экспериментального пика (x, I) имеется соответствующий теоретический $(\tilde{x}, \tilde{I})$. Оценочная функция двух индивидуальных пиков имеет вид $score = s_x * s_I * \sqrt{\tilde{I}}$, где s_x — штраф за расхождение по координате m/z, s_I — штраф за расхождение по интенсивности, а множитель $\sqrt{\tilde{I}}$ введён с целью придавать больший вес высоким пикам. Величины s_x и s_I вычисляются следующим образом:

$$s_x = \max(0, 1 - \frac{|x - \tilde{x}|}{D});$$

$$s_I = \begin{cases} \sqrt{1 - \frac{|I - \tilde{I}|}{I}}, & \tilde{I} < I, \\ 1 - \frac{|I - \tilde{I}|}{I}, & I \leqslant \tilde{I} \leqslant 2I, \\ 0, & \tilde{I} > 2I \end{cases}$$

Здесь D — максимальное расхождение по координате m/z, при котором два пика ещё можно считать совпадающими по ней. MS-Deconv работал с масс-спектрами высокого разрешения и использовал значение $D = 0.02$. Мы же использовали $D = 0.008$. Выбор данного значения согласуется с вычислительными экспериментами, описанными в [6, 7, 8, 9]. Заметим, что штраф за интенсивность более мягкий, если экспериментальный пик оказался выше теоретического, нежели в противном случае. Это связано с тем, что пики двух разных интерферирующих кластеров могут отображаться как один экспериментальный пик, поэтому расхождение в эту сторону не так удивительно.

Вычислив о.ф. для каждой пары пиков, остаётся лишь сложить получившиеся результаты и получить итоговую о.ф. двух масс-спектров.

Опишем теперь, как проводится выбор коэффициента масштабирования t. Величина s_x зависит только от первых координат пиков и не меняется при нормировке. Пусть $\lambda_i = \frac{I}{\tilde{I}^*}$ — отношение интенсивности экспериментального пика с номером i к интенсивности сопоставленного ему теоретического (до масштабирования). В новых терминах множитель s_I можно переписать как

$$s_I = \begin{cases} \sqrt{\frac{t}{\lambda_i}}, & t < \lambda_i \\ 2 - \frac{t}{\lambda_i}, & \lambda_i \leqslant t \leqslant 2\lambda_i, \\ 0, & t > 2\lambda_i \end{cases}$$

Третий множитель будет равен $\sqrt{\tilde{I}} = \sqrt{\frac{tI}{\lambda_i}}$. Несложно проверить, что произведение трёх множителей — функция с единственным максимумом $s_x\sqrt{I}$ в точке $t = \lambda_i$. Эта функция задана кусочно, но на каждом из трёх интервалов имеет простой вид. При изучении суммы нескольких функций такого вида все точки экстремума можно вычислить при помощи дифференцирования. Вычисление о.ф. в каждой из этих точек, однако, слишком затратно по времени, поэтому найти лучший коэффициент масштаба на практике не представляется возможным. Тем не менее, мы можем предложить для него хорошую оценку. Рассмотрим группу теоретических пиков, сопоставленную некоторому экспериментальному пику. Заметим, что составляющие оценочной функции, порождённые этими пиками, имеют одну и ту же точку максимума. Значение о.ф. в этой точке для данной группы пиков несложно вычислить (причём суммарно на это потребуется не больше времени, чем на вычисление о.ф. в любой фиксированной точке). Тогда можно положить $t = \lambda_i$, где i — номер любого из той группы, для которой сумма экстремальных значений о.ф. максимальна. Особо отметим, что на практике в этой точке значение о.ф. близко к теоретическому максимуму, чего нельзя сказать о нормирующем множителе, предложенном в [4].

Фильтрация пиков

Прежде чем переходить к окончательному отбору кандидатов, следует удалить из списка изотопных кластеров часть вариантов, являющихся, скорее всего, ошибочными. Во-первых, мы исключаем из рассмотрения те теоретические кластеры, которым сопоставлено в совокупности менее двух экспериментальных пиков (без учёта фиктивных). Во-вторых, для всех входящих в кластер фиктивных пиков должно существовать общее значение m/z, от которого все фиктивные пики отстоят не более, чем на D. Это условие равносильно тому, что все фиктивные пики находятся внутри интервала ширины $2D$. Также мы для каждого высокого экспериментального пика, интенсивность которого превосходит шумовой порог, оставляем лишь пять изотопных кластеров с сопоставленным ему базовым пиком. Выбор длины списка оставляемых кандидатов равной пяти унаследован от MS-Deconv и обусловлен тем, что шаг отбора алгоритма работает за время, экспоненциальное по длине списка.

Отбор кандидатов

Шаг отбора нашего алгоритма полностью повторяет работу MS-Deconv. На этом шаге мы работаем со списком теоретических изотопных кластеров, для каждого из которых вычислена оценочная функция. Нашей целью является отбор множества кластеров, суммарное значение о.ф. которых будет максимальным. При этом выбранным кластерам запрещено иметь общие пики в экспериментальном масс-спектре. Такую комбинаторную задачу мы сводим к задаче динамического программирования о вычислении пути с наибольшим весом в ориентированном графе без циклов.

Отметим на числовой прямой для каждого кластера $E_i, i = 1, ..., n$, точки s_i и f_i — первые координаты самого левого и самого правого пика кластера. Эти точки разобьют числовую прямую на $2n + 1$ интервал $J_0, J_1, ..., J_{2n}$: два луча и $2n - 1$ отрезков, некоторые из которых, возможно, выродились в точки. Рассмотрим ориентированный граф с вершинами вида (J_k, A), где A — множество кластеров, в котором каждый кластер полностью накрывает отрезок J_k, но сами кластеры не имеют общих пиков. Рёбра в этом графе проводятся из вершин вида (J_k, A) в вершины вида (J_{k+1}, B). Пусть a — общая точка интервалов J_k и J_{k+1}. Тогда мы проводим ребро в графе, если выполнено одно из трёх условий:

- $A = B$;
- $B = A \cup E_i$, $a = s_i$;
- $A = B \cup E_i$, $a = f_i$.

Отметим, что в каждую вершину входит не более одного ребра второго вида. Действительно, эти рёбра входят в те и только те вершины (J_{k+1}, B), для которых левый конец участка J_{k+1} имеет вид s_i для некоторого i, а множество B содержит соответствующий кластер E_i. В таком случае исходящей вершиной ребра будет, очевидно, $(J_k, B \setminus \{E_i\})$. Если в вершину ведёт ребро второго вида, присвоим ей вес $score(E_i)$. В противном случае присвоим ей вес 0. Тогда имеется наглядная биекция между множествами кластеров без общих пиков и путями в графе из вершины $(h_0, \varnothing)$ в вершину $(h_{2n}, \varnothing)$. Более того, сумма весов всех вершин на таком пути равна сумме значений о.ф. тех кластеров, которыми он порождён. Таким образом, для нахождения оптимального набора кластеров (и выделения из них нейтральных масс) достаточно вычислить вес самого "тяжёлого" пути в графе. Это можно сделать, используя метод динамического программирования за время, линейное по числу вершин, которое есть $O(n \cdot 2^l)$, где l — максимальное число кластеров, накрывающих один интервал. Напомним, что для каждого нешумового экспериментального пика мы на шаге обнаружения оставили лишь пять теоретических кластеров с сопоставленным ему базовым пиком, тем самым контролируя значение l.

Результаты

Для оценки качества работы нашего алгоритма мы протестировали его на "триптическом" наборе данных $CAH2$, опубликованном в [12]. Мы работали с файлом 140411_QE_Cah-1.mzXML, содержащим 21880 центрированных масс-спектров, 140411_QE_Cah-1_msdeconv.msalign, содержащим результат работы MS-Deconv на этих масс-спектрах, и 140411_QE_Cah-1.tsv, где некоторые из масс-спектров были идентифицированы при помощи MS-GF+ [10, 11]. Все вычисления проводились на ПК, имеющем процессор с частотой 2,3 ГГц и 8 ГБ оперативной памяти. Для хранения данных и промежуточных результатов работы алгоритма требовалось около 75 МБ свободного места на диске.

В ходе работы с данными мы провели два теста. Целью первого теста было сравнить между собой несколько способов выбора коэффициента масштаба о.ф. В [4] он выбирался так, чтобы сумма интенсивностей самого высокого пика экспериментального кластера и его соседей совпадали с суммой высот соответствующих масштабированных теоретических пиков. Поскольку наш алгоритм учитывает тонкую изотопную структуру, многие экспериментальные пики соответствуют сразу нескольким теоретическим, в связи с чем оригинальное решение невозможно применить без модификаций. Вместо этого в качестве представителя MS-Deconv мы рассмотрели о.ф. $M' - score$, где после масштабирования сумма высот трёх самых интенсивных экспериментальных пиков кластера равна сумме высот трёх самых интенсивных теоретических.

Функцию $M' - score$ мы сравнили с тремя другими о.ф. Функция $S - score$ масштабирует теоретическое распределение так, чтобы его самый высокий пик имел ту же интенсивность, что и самый высокий пик экспериментального изотопного кластера. Функция $U' - score$ выбирает коэффициент масштаба так, как описано в предыдущих разделах. Наконец, функция $T - score$ равна максимальному значению о.ф., которое можно получить при помощи масштабирования.

Чтобы сравнить эффективность разных коэффициентов масштаба, мы выбрали 10 наиболее надёжно идентифицированных масс-спектров. В каждом спектре мы сгенерировали теоретические изотопные кластеры для всех возможных b- и y-ионов идентифицированного пептида и значений заряда от 1 до 3. Чтобы проводить тестирование лишь на тех кластерах, которые действительно представлены в масс-спектре, мы учитывали лишь кластеры, имеющие значение $T - score$ не ниже 250. Средняя доля, которую составляют тестируемые функции от значения $T - score$, указана в таблице.

Таблица 1: Средняя доля о.ф. от максимума в случае правильно определённых кластеров, %

№ скана	$M' - score$	$S - score$	$U' - score$	$T - score$
7153	93,36	92,26	99,11	100
7098	91,35	91,15	99,16	100
14280	97,61	98,29	98,07	100
7208	97,11	95,23	98,85	100
7041	93,36	98,18	99,63	100
13830	96,61	98,02	99,64	100
7528	90,24	91,85	99,32	100
7594	95,08	93,81	97,49	100
13883	99,01	98,96	99,82	100
13780	95,41	98,00	98,99	100

Мы видим, что $U' - score$ превзошёл $M' - score$ во всех случаях, а $S - score$ — во всех случаях, кроме одного. Это позволяет заключить, что именно масштабирование по $U' - score$ приводит к наиболее точному вычислению о.ф. без роста затраченного времени.

Во втором эксперименте мы проверили, как наш алгоритм справляется с деконволюцией тех

масс-спектров, на которых MS-Deconv показывает низкие результаты. Мы отобрали 10 масс-спектров, из которых MS-Deconv выделил не более 10 нейтральных масс, и запустили на них свой алгоритм. Мы называем нейтральную массу попаданием, если она отличается от моноизотопной массы некоторого иона идентифицированного пептида на величину не более $3D$. В противном случае масса называется промахом. Мы изучили число попаданий и промахов среди масс, обнаруженных одновременно нашим алгоритмом и MS-Deconv, а также эти количества среди масс, не выявленных в ходе работы MS-Deconv. Мы заметили, что в некоторых случаях MS-Deconv не выделил ни одной нейтральной массы для кластеров, имеющих высокое значение m/z. В связи с этим мы отдельно изучили число попаданий и промахов среди масс, которые больше, чем все массы, найденные MS-Deconv. Результаты сравнения представлены в таблице.

Таблица 2: Результаты работы алгоритма на масс-спектрах, плохо обработанных MS-Deconv

№ скана	Всего масс	Масс после MS-Deconv	Общие массы	Новые массы	В т.ч. большие
8837	18	10	8/0	5/5	0/0
12113	1	0	0/0	1/0	1/0
11411	20	10	4/6	7/3	4/0
8773	13	8	6/1	5/1	1/0
9227	12	5	2/1	6/3	4/2
9042	10	8	5/2	1/2	0/0
9289	10	3	1/1	6/2	3/1
8970	11	9	6/1	3/1	0/0
9271	20	5	5/0	8/7	5/3
8309	6	7	3/1	1/1	0/0

В столбцах 4-6 первое число соответствует количеству попаданий среди рассматриваемых масс, а второе число — количеству промахов.

Мы видим, что в 9 случаях из 10 наш алгоритм находит больше масс, чем MS-Deconv, при этом доля попаданий среди выявленных масс сопоставима с той, которая получается в результате работы MS-Deconv. Так мы можем заключить, что новый алгоритм показывает более хорошие результаты при работе с некачественными масс-спектрами. Наличие масс, успешно обнаруженных MS-Deconv, но не вошедших в итоговый список при запуске нашего алгоритма, вероятно, обусловлено выбором MS-Deconv более низкого уровня шума.

Заключение

В этой работе мы предложили альтернативу алгоритму MS-Deconv, которую можно впоследствии применять для анализа bottom-up масс-спектров высокого и сверхвысокого разрешения. В дальнейшем планируется объединить все используемые программы в единый пайплайн, а так же продолжить тестирование алгоритма на данных, снятых по другим технологиям. Для представленного алгоритма можно предложить несколько дальнейших улучшений. Во-первых, качество результатов может возрасти, если уровень шума выбирать динамически (например, регулируя долю нешумовых пиков в масс-спектре). Во-вторых, потенциально могут быть улучшены методы фильтрации пиков. Для bottom-up масс-спектров нередки изотопные кластеры, состоящие из небольшого числа пиков, и их следует анализировать иначе, нежели более крупные. В-третьих, время работы шага обнаружения кандидатов можно уменьшить, если заранее сгенерировать базу данных изотопных кластеров. В-четвёртых, алгоритм можно распространить на случай, когда в масс-спектр входят ионы типов, отличных от b и y. Наконец, интересно найти аналог задачи отбора изотопных кластеров, в котором различные взятые кластеры могут иметь общие пики. Все эти улучшения будут целью последующих исследований.

Список литературы

[1] N. L. Kelleher, H. Y. Lin, G. A. Valaskovic, G. A. Aaserud, E. K. Fridriksson, F. W. McLafferty, Top down versus bottom up protein characterization by tandem highresolution mass spectrometry. *Journal of American Chemical Society*, **121**:4 (1999), 806–812.

[2] B.T. Chait, Mass spectrometry: Bottom-up or top-down? *Science*, **314**:5796 (2006), 65–66.

[3] Q. Hu, H. Li, A. Makarov, M. Hardman, R. G. Cooks, The Orbitrap: A new mass spectrometer. *Journal of Mass Spectrometry*, **40**:4 (2005), 430–433.

[4] X. Liu, Y. Inbar, P. C. Dorrestein, C. Wynne, N. Edwards, P. Souda, J. P. Whitelegge, V. Bafna, P. A. Pevzner, Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins. *Molecular & Cellular Proteomics*, **9**:12 (2010), 2772–2882.

[5] A. L. Rockwood, P. Haimi, Efficient Calculation of Accurate Masses of Isotopic Peaks. *Journal of American Society for Mass Spectrometry*, **17** (2006), 415–419.

[6] K. Vyatkina, S. Wu, L. J. M. Dekker, M. M. VanDuijn, X. Liu., N. Tolić, M. Dvorkin, S. Alexandrova, T. M. Luider, L. Paša-Tolić, P.A. Pevzner, De novo sequencing of peptides from top-down tandem mass spectra. *Journal of Proteome Research*, **14**:11 (2015), 4450–4462.

[7] K. Vyatkina, S. Wu, L. J. M. Dekker, M. M. VanDuijn, X. Liu., N. Tolić, T. M. Luider, L. Paša-Tolić, P.A. Pevzner, Top-down analysis of protein samples by de novo sequencing techniques. *Bioinformatics*, **32**:18 (2016), 2753–2759.

[8] K. Vyatkina, De novo sequencing of top-down tandem mass spectra: A next step towards retrieving a complete protein sequence. *Proteomes*, **5**:1 (2017), 6.

[9] K. Vyatkina, L. J. M. Dekker, S. Wu, M. M. VanDuijn, X. Liu., N. Tolić, T. M. Luider, L. Paša-Tolić, P.A. Pevzner, De novo sequencing of peptides from high-resolution bottom-up tandem mass spectra using top-down intended methods. *Proteomics*, **17**:23–24 (2017).

[10] S. Kim, N. Gupta, P. Pevzner, Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *Journal of Proteome Research*, **7**:8 (2008), 3354-3363.

[11] S. Kim, P. Pevzner, MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, **5** (2014), 5277.

[12] K. A. Cupp-Sutton, S. Wu, High-throughput quantitative top-down proteomics. *Molecular Omics*, **16**:2 (2020), 91-99