Санкт-Петербургский государственный университет

ФИЛИППОВ Степан Дмитриевич

Выпускная квалификационная работа

Алгоритмы анализа белковых последовательностей с использованием внутренних фрагментных ионов

Уровень образования: бакалавриат
Направление 01.03.01 «Математика»
Основная образовательная программа СВ.5000.2019 «Математика»

Научный руководитель:
доцент факультета математики и компьютерных наук,
д.ф.-м.н. Степанов Алексей Владимирович

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

Санкт-Петербург
2023 год

Содержание

1 Введение **1**

2 Постановка задачи **1**

3 Обозначения **2**

4 Описание и подготовка данных **2**

5 Проверка статистических гипотез **2**

6 Подход с использованием *сырых* спектров **3**

7 Подход с использованием деконволюцированных спектров **5**
7.1 Начальные и конечные ионы 5
7.2 Внутренние ионы 7

8 Подход с использованием *лесенок* **8**
8.1 Определение *лесенок* 8
8.2 Алгоритм нахождение *лесенок* 9
8.3 Примеры найденных *лесенок* 10
8.4 Анализ позиций *разломов* 12

9 Заключение **12**

10 Исходный код решения **13**

1 Введение

Масс-спектрометрия – это метод, позволяющий идентифицировать и анализировать химические вещества. Изучаемые молекулы ионизируются и расщепляются на части, после чего масс-спектрометр измеряет отношение массы к заряду различных компонент и объем их присутствия в образце с помощью измерения интенсивности ионного тока.

В протеомике масс-спектрометрия используется для идентификации белков и пептидов. Существует два основных подхода, *top-down* и *bottom-up* [1, 2]. В случае *bottom-up* подхода, белок предварительно расщепляется на короткие пептиды, *top-down* подход использует весь белок целиком.

В работах [3, 4, 5, 6] был представлен алгоритм Twister, решающий задачу *de novo* секвенирования белков, то есть идентификации последовательности аминокислот в белке на основе набора *top-down* масс-спектров. Данный алгоритм опирается на тот факт, что фрагментные ионы, соответствующие пикам в масс-спектре, часто образуют последовательности расширяющихся фрагментов с общим концом. Алгоритм в первую очередь опирается на ионы, являющиеся началом или концом всего белка, но в масс-спектрах бывают также и внутренние ионы. Информация об их устройстве и распределении может помочь усовершенствовать алгоритмы *de novo* секвенирования белков.

В данной работе предполагается проанализировать поведение внутренних фрагментных ионов в масс-спектрах.

2 Постановка задачи

Масс-спектр представляет из себя набор пиков, состоящих из величины отношения массы к заряду и интенсивности. При проведении исследований получается множество масс-спектров, не все из них являются масс-спектром для всего белка, часть из них соответствуют префиксу, суффиксу или инфиксу белка. Часть масс-спектров могут соответствовать посторонним молекулам, случайно попавшим в смесь. Но информация о том, чему соответствует каждый масс-спектр неизвестна и должна быть получена алгоритмическими путями. Задачей является исследование *top-down* масс-спектров для известного белка CAH2. Для каждого масс-спектра требуется понять, из какого фрагмента белка он был получен, а также разметить пики в этом масс-спектре. Пики могут быть отнесены к начальным, конечным, а также внутренним фрагментам ионов. Кроме того, необходимо проанализировать полученные результаты и попытаться найти закономерности в том, в каких позициях последовательности аминокислот чаще всего начинаются или заканчиваются внутренние ионы.

3 Обозначения

Symbol	Definition
s	исследуемый белок, CAH2
$s[l:r]$	фрагмент белка с позиции l до r
$m(s)$	масса последовательности s (Da)
length(s)	длина последовательности s
H_2O	масса молекулы воды (Da)
NH_3	масса молекулы аммиака (Da)
p	масса протона (Da)

Таблица 1: Обозначения

4 Описание и подготовка данных

Рассматривались масс-спектры, снятые с белка CAH2_BOVIN. В подходе с использованием *сырых* масс-спектров использовались MS/MS спектры. Подробнее информация о данном наборе масс-спектров приведена в статье [6].

Алгоритмы, предназначенные для работы с деконволюцированными масс-спектрами, были протестированы на масс-спектрах, которые были получены из выше описанного набора масс-спектров с помощью алгоритма деконволюции MS-Deconv [7]. Подробное описание параметров, использованных для запуска алгоритма MS-Deconv, можно также найти в статье [6].

5 Проверка статистических гипотез

В разных местах в данной работе мы будем в масс-спектрах находить объекты (аннотированные пики, *лесенки*), однако всякий раз будет возникать вопрос: данный объект на самом деле существует, или мы случайно нашли некую закономерность в шуме? Мы будем считать статистики и проверять статистические гипотезы, чтобы в каждом случае принимать решение о том, стоит ли обращать внимание на найденный объект.

Напомним, как работает проверка статистических гипотез. Мы рассматриваем имеющиеся данные D как результаты статистического эксперимента. Во всех случаях в данной работе нулевая гипотеза будет простой, то есть состоять из одного распределения X_0. Обычно желаемым результатом проверки статистической гипотезы является отклонение нулевой гипотезы. Выбирается статистика τ и считается значение статистики для имеющихся данных $\tau(D)$, а так же для большого числа синтетических данных $\{D_1, D_2, \ldots, D_N\}$, сгенерированных из X_0. После этого считается p-value, а именно доля синтетических данных, на которых значение статистики не меньше $\tau(D)$. После этого полученное p-value сравнивается с уровнем значимости (мы используем 0.05 или 0.1).

Рассмотрим немного подробнее, как именно будут проверяться статистические гипотезы в случае с анализом белковых последовательностей. Пусть есть исследуемый

белок s, рассматривается масс-спектр m, и на их основе некоторый алгоритм $A(s, m)$ выдает какой-то объект, например, множество аннотированных пиков в масс-спектре. Обычно такой алгоритм не может давать гарантии, что найденный объект действительно существует, а не был случайно найден в шуме.

Нулевая гипотеза будет отражать то, что данное множество пиков было аннотировано случайно. Для этого рассмотрим распределение X_0 на всех последовательностях аминокислот длины $length(s)$. На каждой позиции аминокислоты выбирается равновероятно среди 20 стандартных аминокислот. Кроме того, на разных позициях аминокислоты выбираются независимо. Статистикой $\tau(s)$ будет размер множества $size(A(s, m))$. Идея, стоящая за этим, заключается в том, что большой размер множества $A(s, m)$ дает больше уверенности в том, что найденный объект существует на самом деле, так как очень маловероятно случайно аннотировать большое число пиков. Поэтому сгенерируем N случайных последовательностей аминокислот $s_1, s_2, \ldots, s_N$ из распределения X_0, посчитаем величины $size(A(s_1, m)), \ldots, size(A(s_N, m))$ и определим долю величин из этого списка, не меньших $size(A(s, m))$. Полученное p-value сравниваем с уровнем значимости α. Альтернативно, можно посчитать $1-\alpha$ квантиль данного списка значений и проверить, что $size(A(s, m))$ не меньше этого квантиля. В любом случае, если проверка успешна, то мы отвергаем нулевую гипотезу, что дает нам уверенность в том, что мы действительно обнаружили закономерность в данных.

Впоследствии мы не будем каждый раз подробно описывать процедуру проверки статистической гипотезы, а будем только указывать нулевую гипотезу X_0, статистику τ, число N и уровень значимости α.

6 Подход с использованием *сырых* спектров

Рассмотрим алгоритмы работы с *сырыми* масс-спектрами, то есть случай, когда каждый пик характеризуется отношением массы к заряду (m/z).

Пусть молекула, имеющая массу m, попадает в масс-спектрометр. Масс-спектрометр может детектировать только заряженные частицы, поэтому сначала на него помещаются z протонов, сообщающих ему заряд $+z$. После этого масса иона становится равной $m+z\cdot p$, где p обозначает массу протона ($p = 1.00727646677$ Da). Масс-спектрометр измеряет отношение массы к заряду, которое мы будет обозначать как mz, которое выражается через массу и заряд как $mz = \frac{m+z\cdot p}{z} = \frac{m}{z} + p$. Преобразуя это выражение, получаем равенство

$$m = (mz - p) \cdot z$$

Так как заряд z неизвестен, то нужно пробовать все возможные варианты в некоторых пределах. В данной работе предполагалось, что $1 \leq z \leq 50$, исходя из того, что масса всего белка составляет примерно 29006.68 Da, а на 1000 Da в среднем редко приходится больше одного заряда.

Мы будем использовать алгоритм 1 как для определения массы прекурсора, так и для аннотации b-ионов и y-ионов. Алгоритм принимает на вход два параметра: mz и $candidates$. Параметр mz - это значение m/z иона, который нас интересует. Параметр $candidates$ - это список нейтральных молекул, среди которых может находиться молекула, из массы которой было получено интересующее нас значение mz.

Сначала алгоритм инициализирует пустой список $explanations$, в который мы будем собирать все возможные варианты получить значение mz из какой-то из масс

среди *candidates*.

Для каждого значения массы $m(i)$ молекулы i из списка *candidates* мы перебираем заряды в диапазоне от 1 до 50. Для каждой пары $m(i)$ и z вычисляется теоретическое значение m/z, обозначенное как t_mz. После этого теоретическое значние t_mz сравнивается с mz с точностью 10 ppm. Все найденные варианты добавляются в список *explanations*, который в итоге возвращается из функции.

Algorithm 1 Deconvolute and Annotate

```
1: Global Variables:
2: p ← mass of a proton
3: err ← 10 ppm
4:
5: Input: mz, candidates
6: Output: explanations
7:
8: procedure DeconvoluteAndAnnotate(mz, candidates)
9:     explanations ← empty list
10:    for each molecule i in candidates do
11:        for charge z from 1 to 50 do
12:            t_mz ← (m(i) + z · p)/p
13:            if t_mz is within mz ± (mz · err/1e6) then
14:                Append (i, z) to explanations
15:            end if
16:        end for
17:    end for
18:    return explanations
19: end procedure
```

Данный алгоритм использовался в двух случаях. Во-первых, с помощью него можно попытаться найти масс-спектры, в которых прекурсором является вся молекула, а не протеолитический фрагмент. Для этого в функцию *DeconvoluteAndAnnotate* можно передать m/z прекурсора, а в качестве *candidates* весь белок. Используя более качественные алгоритмы, о которых речь пойдет дальше, результат этого алгоритма был проверен. Результаты совершенно неудовлетворительные: алгоритм не только не нашел много масс-спектров, снятых со всего белка, но и среди найденных масс-спектров большинство являются ложными срабатываниями. Первое скорее всего объясняется тем, что когда прекурсором является очень длинная последовательность аминокислот, в ней наверняка происходит некоторое количество модификаций, которые приводят к изменению массы. Вторую проблему данного алгоритма лучше иллюстрирует второе применение его.

Мы также пробовали использовать данный алгоритм для аннотирования пиков в масс-спектрах. Для этого в функцию *DeconvoluteAndAnnotate* передавалось значение m/z для интересующего нас пика, а также список всех *b*-ионов и *y*-ионов. Кроме того, так как ионы часто могут терять молекулу воды или аммиака, то вместе с ионом i массы $m(i)$ в список добавляются модифицированные ионы с массами $m(i)-m(H_2O)$ и $m(i) - m(NH_3)$.

Для каждого пика в масс-спектре запускается данный алгоритм. Как уже было

сказано выше, этот алгоритм имеет много ложных срабатываний. Это происходит из-за того, что мы перебираем 50 возможных значений заряда z. Поэтому для каждого пика мы делаем очень много сравнений, и есть немаленькая вероятность, что масса данного пика случайно окажется в пределах 10 ppm от величины t_mz для хотя бы одной пары значений $m(i)$ и z.

Чтобы проверить это строго, мы использовали статистический тест. В данном случае статистический тест проводился не для отдельного пика, а для всего масс-спектра. А именно, сначала алгоритм 1 запускался на всех пиках масс-спектра и пытался соотнести их с теоретическими массами b-ионов и y-ионов исследуемого белка. Далее считалась статистика τ, равная количеству пиков, которые удалось аннотировать. После этого данная процедура повторялась $N = 100$ раз, но в качестве $candidate\_masses$ использовались массы ионов случайно сгенерированной последовательности белков. Потом считалось p-value, то есть доля искусственных белков, на которых значение статистики не меньше значения статистики на исследуемом белке. Во всех случаях p-value получалось больше 0.05 и нулевая гипотеза не отклонялась. Это означает, что мы не могли сделать вывод о том, что аннотированные пики действительно соответствуют ионам исследуемого белка, а не были соотнесены с каким-то ионом случайно.

В целом, подход с использованием *сырых* масс-спектров себя не оправдал, поэтому дальше мы перешли к работе с деконволюцированными масс-спектрами.

7 Подход с использованием деконволюцированных спектров

7.1 Начальные и конечные ионы

Сырые MS/MS спектры были деконволюцированы с помощью алгоритма MS-Deconv [7]. Это означает, что теперь каждый пик в масс-спектре имеет нейтральную массу, и не нужно перебирать возможные значения заряда. Это сильно снижает количество вариантов, которые нужно проверять при соотнесении пиков с теоретическими массами ионов, так что есть надежда, что теперь получится создать алгоритм с меньшим числом ложных срабатываний.

Мы будем использовать алгоритм 2, похожий на предыдущий, но в нем не нужно будет перебирать варианты заряда.

Применим новый алгоритм для аннотирования пиков в MS/MS спектрах. Для этого сгенерируем список *candidates*, состоящий из всех b-ионов и y-ионов. Как и в прошлый раз, включаем также модификации ионов, получающиеся потерей воды или аммиака. Будем передавать в функцию *Annotate* массы пиков и смотреть, какие из них удалось соотнести с каким-то из теоретических ионов.

Как и в предыдущий раз, будем использовать статистический тест для оценки нашей уверенности в аннотированных ионах. Снова запускаем всю процедуру для $N = 100$ случайно сгенерированных белков, а в качестве статистики используем число аннотированных пиков в деконволюцированном масс-спектре. Посчитаем 0.95-квантиль распределения статистики на случайно сгенерированных белках. В таблице 2 приведена информация о числе аннотированных пиков у некоторого набора масс-спектров, а также информация о том, сколько пиков получается аннотировать, если соотносить их с теоретическими массами случайных белков. Как мы видим, количе-

Algorithm 2 Annotate

```
Global Variable:
err ← 10 ppm

Input: m, candidates
Output: explanations

procedure ANNOTATE(m, candidates)
    explanations ← empty list
    for each molecule i in candidates do
        if m(i) is within m ± (m · err/1e6) then
            Append i to explanations
        end if
    end for
    return explanations
end procedure
```

ство аннотированных пиков в случае случайных белков гораздо меньше, что не только отклоняет нулевую гипотезу, но и позволяет количественно оценить примерную долю случайно аннотированных пиков. Глядя на таблицу, можно сделать вывод о том, что не больше 10% аннотированных пиков были случайно соотнесены с некоторыми ионами, поэтому можно с большой степенью уверенности заявлять, что большинство аннотированных пиков действительно соответствуют b-ионам и y-ионам.

номер	**число пиков**	**число аннотированных пиков**	0.95 **квантиль**
839	69	30	2
1074	75	31	2
1288	150	34	4
2309	150	54	3
2314	467	100	10
2408	98	17	2
2479	241	30	4

Таблица 2: Результаты аннотирования для некоторых MS/MS спектров

Кроме того, можно посмотреть на распределение позиций, в которых чаще всего происходит "разлом то есть позиции, в которых часто заканчиваются или начинаются найденные в масс-спектре ионы. Последовательность аминокислот, образующая молекулу CAH2, имеет длину 259, так что мы можем пронумеровать позицию перед первой аминокислотой, позиции между последовательными аминокислотами и позицию после последней аминокислоты числами $0, 1, \ldots, 259$. Будем откладывать позиции по горизонтальной оси, а по вертикальной оси будем откладывать суммарное по всем MS/MS спектрам количество аннотированных b-ионов, которые заканчиваются в данной позиции (синий цвет), и суммарное количество y-ионов (красный цвет). Результат изображен на рисунке 1.

Замечание. Позиции, в которых часто заканчиваются или начинаются фрагментарные ионы, распределены очень неравномерно. Есть небольшое количество мест, на

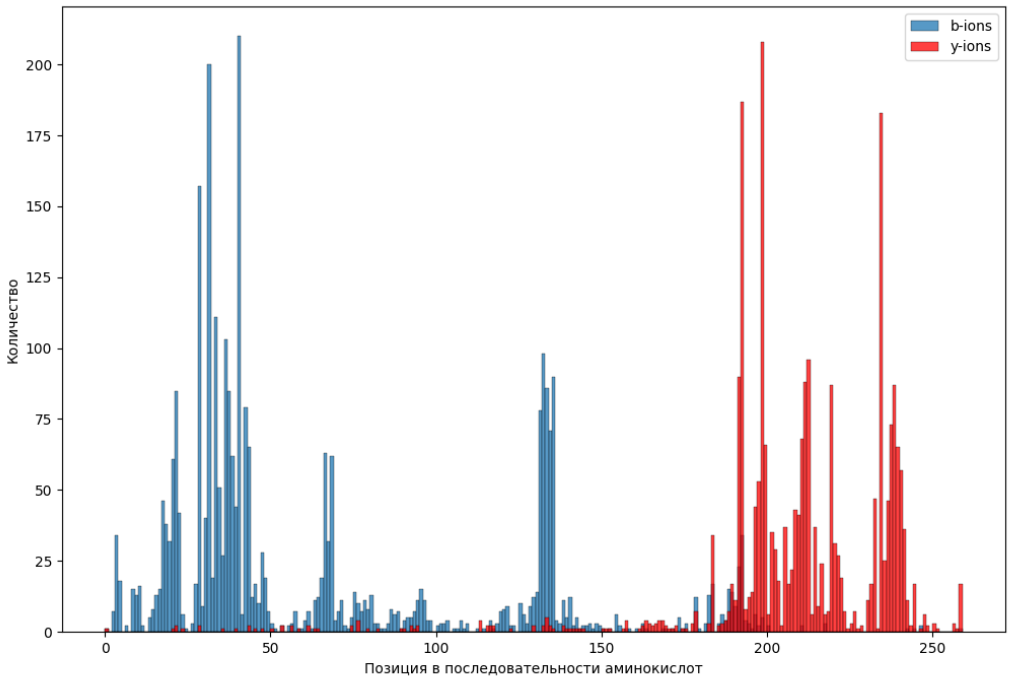

Рис. 1: Гистограмма, отбражающая как часто каждая позиция в последовательности аминокислот белка CAH2 является концом b-иона (синий) или началом y-иона (красный)

которые приходится большинство *разломов*. Далее в данной работе мы более внимательно рассмотрим такие позиции.

7.2 Внутренние ионы

После успеха в нахождении начальных и конечных ионов в масс-спектрах, перейдем к нахождению внутренних ионов. Внутренним ионом называется фрагмент прекурсора, который не является ни началом, ни концом, то есть не является b-ионом или y-ионом.

Мы можем повторить всю процедуру для поиска внутренних ионов. Для этого расширим список *candidates* всеми возможными внутренними фрагментами исследуемого белка.

После этого повторим статистический тест, такой же, как для начальных и конечных ионов. В таблице 3 приведена информация о числе аннотированных пиков для исследуемого белка, а также 0.95-квантиль распределения числа аннотированных белков, если использовать внутренние ионы случайно сгенерированной последовательности аминокислот.

К сожалению, в некоторых случаях вполне вероятна ситуация, при которой в случайном белке аннотируется больше пиков, чем в истинном. Но даже если это не так, то все равно отличить аннотированные случайно от ионов, которые действительно присутствуют в масс-спектре данный алгоритм не позволяет, поэтому данный подход

номер	**число пиков**	**число аннотированных пиков**	0.95 **квантиль**
839	69	23	39
1074	75	31	43
1288	150	83	89
2309	150	74	49
2314	467	226	209
2408	98	34	25
2479	241	105	92

Таблица 3: Результаты аннотирования внутренних ионов для некоторых MS/MS спектров

не может быть применим для нахождения внутренних ионов.

Как и прежде, проблема ложных результатов возникает из-за того, что внутренних ионов гораздо больше, чем начальных и конечных. Поэтому для каждого пика мы выполняем больше сравнений, что повышает вероятность того, что масса пика случайно окажется похожей на теоретическую массу иона.

Необходимо придумать способ понимать достоверность найденных внутренних ионов. Для этого мы рассмотрим понятие *лесенки* и предложим алгоритм, который позволяет находить внутренние ионы, в которых есть уверенность, что они не случайно совпали с массами пиков.

8 Подход с использованием *лесенок*

8.1 Определение *лесенок*

Как уже было замечено в предыдущих разделах, места *разломов* белка при масс-спектрометрии распределены внутри последовательности аминокислот неравномерно. Есть позиции, в которых ионы начинаются или заканчиваются очень часто, в других позициях *разломы* происходят очень редко или вообще не происходят.

Отдельный интерес вызывают наборы ионов, имеющие общий левый или правый конец. Поэтому введем соответствующее определение.

Определение. *Лесенкой* будет называться набор ионов (фрагментов исследуемого белка), имеющих общий левый или правый конец. В случае, когда важно уточнить, какой конец является общим, будем использовать понятие левая *лесенка* или правая *лесенка*.

Ценность *лесенок* заключается в том, что именно они позволяют алгоритмам, таким как Twister, получать наборы k-тагов и секвенировать белки на основе top-down MS/MS спектров.

Например, левой *лесенкой* является набор из всех аннотированных b-ионов, а правой *лесенкой* является набор из аннотированных y-ионов. Кроме того, возможны ситуации, когда в масс-спектре присутствует некоторое количество пиков, которые соответствуют внутренним ионам, начинающимся или заканчивающимся в одной позиции.

Нахождение *лесенок* из внутренних ионов и их изучение является основной задачей данной работы. В связи с этим и тем фактом, что именно внутренние ионы, образующие *лесенки*, позволяют успешно секвенировать белки, мы сосредоточим наши усилия на поиске именно тех ионов в масс-спектрах, которые образуют *лесенки*.

8.2 Алгоритм нахождение *лесенок*

Перейдем теперь к алгоритму 3, который позволит находить *лесенки* внутри MS/MS спектров. Основная идея заключается в том, что если аннотировать пики в масс-спектре не по отдельности, а группами, образующими *лесенки*, то такие аннотированные ионы являются гораздо более достоверными. Действительно, хотя есть немалая вероятность, что произвольная масса случайно окажется достаточно похожей на массу одного из внутренних ионов, вероятность того, что много пиков будут случайно соотнесены с ионами с общим началом, гораздо меньше. Мы проверим это с помощью статистического теста, а пока перейдем к описанию алгоритма.

На вход алгоритм получает m_array – список масс пиков MS/MS спектра и $ions$ – список всех b-ионов исследуемого белка. Важно, что в $ions$ мы не включаем внутренние ионы. b-ионы используются для нахождения левых *лесенок*, чтобы найти правые *лесенки*, нужно запустить этот же алгоритм еще раз, но в аргумент $ions$ передать список всех y-ионов.

В алгоритме используются процедуры $CreatePotentialPairs$ и $GroupSegments$. Первая процедура принимает набор масс m_array и набор ионов $ions$, и для каждой пары, состоящей из массы m и иона i, считает значение $m(i) - m$. Кроме того, чтобы учесть возможную потерю или наоборот присоединение молекулы воды или аммиака, добавляются модификации.

Идея считать все возможные разности между массами из MS/MS спектра и теоретическими массами ионов заключается в следующем. Предположим, что в масс-спектре есть левая *лесенка* размера n, состоящая из масс $\{m_i\}_{i=1}^n$. Пусть соответствующие ионы имеют общий левый конец, равный l, а правые концы разные и равны $\{r_i\}_{i=1}^n$. Нетрудно видеть, что выражения $m(s[:r_i]) - m_i$, с точностью до погрешности измерения, равны одному и тому же числу, $m(s[:l])$.

Так как мы не знаем значения l и r_i, то мы пробуем все возможные варианты правых концов. В случае с правильными значениями r_i мы получим много очень похожих масс, которые все будут приближенно равны $m(s[:l])$. Так что следующий шаг заключается в том, чтобы в большом наборе попарных разностей выделить группы близких значений.

Для этого используется процедура $GroupSegments$. Поданный на вход массив сортируется по возрастанию, после чего считается, что если соседние элементы отличаются не больше, чем на $tol = 4 \cdot 10^{-3}$ (подобрано эмпирически), то эти элементы должны быть отнесены в одну группу. После этого с помощью одного прохода по массиву все элементы разбиваются на группы.

Наконец, процедура выделения лесенок в масс-спектре работает следующим образом. Используется жадный алгоритм. На каждой итерации сначала вызывается $CreatePotentialPairs$, чтобы посчитать массив разностей, а затем $GroupSegments$ разбивает массив разностей на группы. Затем выбирается самая большая группа, которая становится потенциальной *лесенкой*. Если ее размер меньше 5, то алгоритм завершает работу. В противном случае, *лесенка* добавляется в список *лесенок*, а из

масс-спектра удаляются массы, вошедшие в новую лесенку. После этого алгоритм переходит на следующую итерацию и заново считает массив разностей.

Но это еще не конец. Дело в том, что даже в этом алгоритме выделяется некоторое количество (не очень длинных) лесенок, которые возникают случайным образом. Это также было проверено с помощью симуляции работы алгоритма на случайно сгенерированных последовательностях аминокислот. А именно, генерировалось $N = 100$ случайных последовательностей аминокислот (той же длины, что и исследуемый белок), и для каждой из них с помощью алгоритма 3 находилась длина наибольшей лесенки.

Изначально данная статистическая информация использовалась для того, чтобы отфильтровать слишком короткие *лесенки*, в которых у нас нет уверенности, что они являются настоящими. А именно, находился 0.95-й квантиль распределения длины наибольшей лесенки у случайного белка, и полученное значение использовалось в качестве порога для фильтрации *лесенок*. То есть оставлялись только те *лесенки*, которые были длиннее порогового значения.

После этого было замечено, что мы не используем информацию об общем левом конце найденных *лесенок*. Мы знаем, что $m(s[:\, l]) = m(s[:\, r_i]) - m_i$, поэтому зная r_i и m_i для найденной *лесенки*, мы можем вычислить приближенную массу $m(s[:\, l])$ и затем восстановить значение l. Для этого мы записываем теоретические массы всех b-ионов и находим ближайшее к нашему значению. Длина соответствующего иона будет равна l. После этого мы оставляем только те *лесенки*, у которых $m(s[:\, l])$ не отличается от $m(s[:\, r_i]) - m_i$ или отличается на $\{+H_2O, -H_2O, +NH_3, -NH_3\}$. Все сравнения проводятся с точностью до 0.03 Da. Затем мы повторяем всю эту процедуру для рандомных последовательностей аминокислот и снова вычисляем 0.95-й квантиль распределения длины наибольшей *лесенки*. Но теперь мы фильтруем большинство ложных *лесенок* на стадии вычисления левой границы, поэтому в данном случае 0.95-й квантиль получается гораздо меньше, что позволяет нам с большой долей уверенности выделять в том числе и короткие *лесенки*.

8.3 Примеры найденных *лесенок*

На рисунке 2 изображены 4 *лесенки*, которые были найдены в одном MS/MS спектре. Верхние две *лесенки* составлены из b-ионов и y-ионов соответственно, а нижние - из внутренних ионов.

Информация о найденных *лесенках* во многих случаях позволяет также определить, каким фрагментом всего белка является прекурсор, с которого был снят данный MS/MS спектр. Дело в том, что *лесенки* из b-ионов и y-ионов встречаются почти во всех масс-спектрах, поэтому можно рассмотреть самое левое начало среди левых *лесенок* и самый правый конец среди правых *лесенок* - это будет хороший кандидат на роль прекурсора. Для большей уверенности можно сравнить теоретическую массу получившегося фрагмента с массой прекурсора, посчитанной алгоритмом MS-Deconv.

Например, в случае рисунка 2 можно сделать вывод, что прекурсором является вся молекула CAH2 целиком, а в случае рисунка 3 прекурсором скорее всего является протеолитический фрагмент, являющийся началом CAH2.

Algorithm 3 Generate All Ladders

```
procedure GenerateAllLadders(m_array, ions)
    ladders ← empty list
    tol ← 4 · 10^−3

    procedure CreatePotentialPairs(m_array, ions)
        result ← empty list
        modifications ← {0, +H2O, −H2O, +NH3, −NH3, +1, −1}
        for each m in m_array do
            for each ion i in ions do
                for each mod in modifications do
                    start_pos ← m(i) − m + mod
                    if start_pos ≥ −0.01 then
                        Add (start_pos, m, i) to result
                    end if
                end for
            end for
        end for
        return result
    end procedure

    procedure GroupSegments(pairs_list)
        groups ← empty list
        Sort pairs_list by start_pos
        current_group ← empty list
        for i from 0 to length of pairs_list do
            if current_group is empty or distance between starting positions for consecutive elements ≤ tol then
                Add pairs_list[i] to current_group
            else
                Add current_group to groups
                current_group ← empty list
            end if
        end for
        return groups
    end procedure

    while True do
        pairs_list ← CreatePotentialPairs(m_array, ions)
        groups ← GroupSegments(pairs_list)
        largest_ladder ← maximal group in groups
        if length of largest_ladder < 5 then
            break
        end if
        Add largest_ladder to ladders
        Remove masses that form largest_ladder from m_array
    end while
    return ladders
end procedure
```

8.4 Анализ позиций *разломов*

Ранее мы уже отмечали неравномерное распределение позиций в последовательности аминокислот, в которых часто начинается или заканчивается фрагментный ион. Теперь мы более внимательно посмотрим на эти позиции.

Во-первых, ориентируясь на найденные *лесенки* и массу прекурсора для большинства MS/MS спектров, определим фрагмент, с которого был снят этот спектр. Это может быть весь белок или протеолитический фрагмент. После этого все *лесенки* можно разделить на внутренние, то есть полностью содержащиеся внутри прекурсора, и граничные, то есть те, у которых общий конец совпадает с границей прекурсора.

Сконцентрируем внимание на внутренних *лесенках*. Как и все *лесенки*, они делятся на левые и правые в зависимости от того, какой из концов ионов общий. Можно предположить, что на вероятность *разлома* в конкретной позиции может влиять то, какие аминокислоты находятся сразу перед или после места разлома.

P	D	A	M	K	H	Q	V
63	2	2	1	1	1	1	1

Таблица 4: Количество раз, когда внутренняя левая *лесенка* начинается с данной аминокистоты

D	L	K	A	G	N	E	Q
10	2	1	1	1	1	1	1

Таблица 5: Количество раз, когда внутренняя правая *лесенка* заканчивается данной аминокистотой

Замечание. В случае молекулы CAH2 есть ярко выраженная закономерность в том, в каких местах начинаются и заканчиваются лесенки из внутренних фрагментных ионов. А именно, абсолютное большинство из найденных левых *лесенок* начинаются с пролина (P), а большинство найденных правых *лесенок* заканчиваются аспарагиновой кислотой (D).

9 Заключение

В этой работе была рассмотрена задача аннотирования ионов, в первую очередь внутренних ионов, белковых последовательностей. Последовательными улучшениями алгоритма удалось достоверно находить составленные из внутренних ионов *лесенки* в масс-спектрах. В свою очередь, это позволило определять, с какого фрагмента белка был снят масс-спектр, а также сделать некоторые выводы о том, как устроены места, в которых происходит расщепление белка на фрагменты.

10 Исходный код решения

Все алгоритмы были реализованы на языке программирования Python с использованием библиотеки pyteomics. Код алгоритмов доступен по ссылке:

https://github.com/pingpong17/MS-MS-spectrum-analysis

Список литературы

[1] Neil L. Kelleher, Hong Y. Lin, Gary A. Valaskovic, David J. Aaserud, Einar K. Fridriksson, and Fred W. McLafferty, *Top Down versus Bottom Up Protein Characterization by Tandem High-Resolution Mass Spectrometry*, Journal of the American Chemical Society 1999 **121** (4), 806-812, DOI: 10.1021/ja973655h

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

[3] 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.

[4] 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.

[5] 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.

[6] 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).

[7] 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 2782.

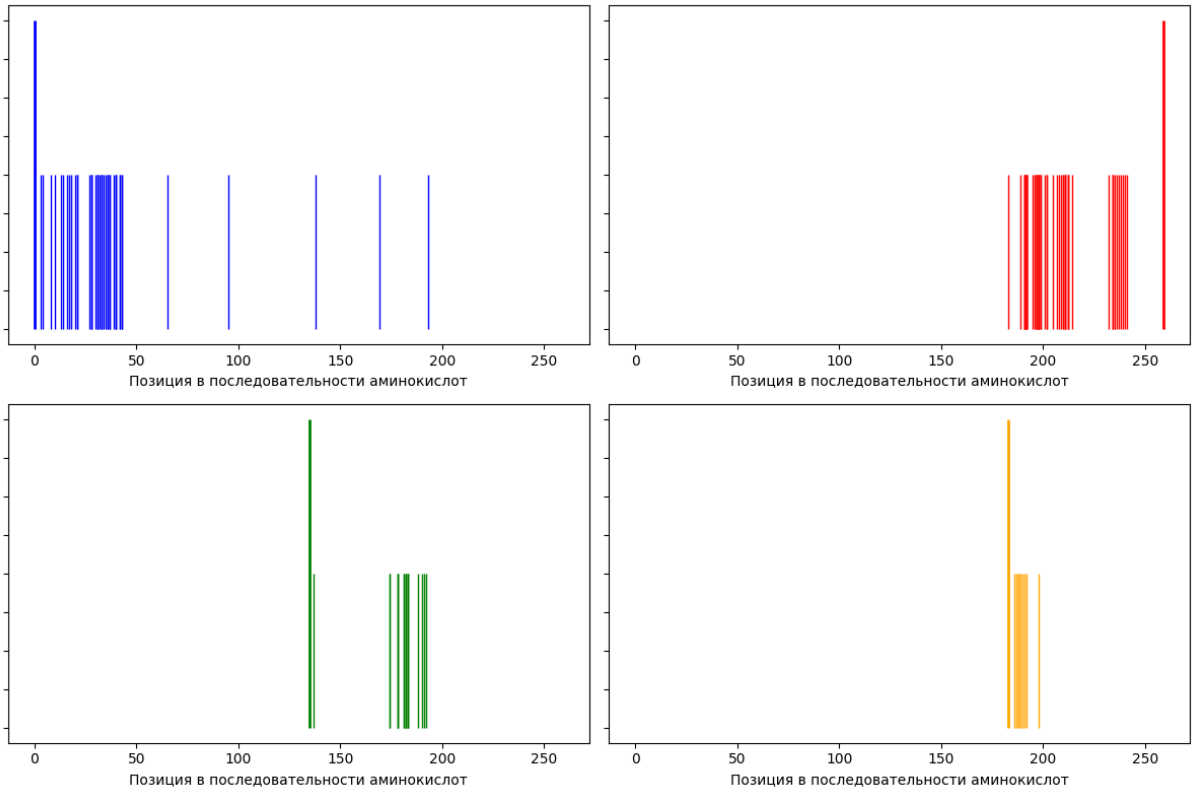

Рис. 2: В левом верхнем углу изображена *лесенка* из b-ионов, в правом верхнем угла *лесенка* из y-ионов, снизу изображены две *лесенки* из внутренних ионов. Высокая линия обозначает общее начало *лесенки*, короткие линии обозначают концы ионов в *лесенке*.

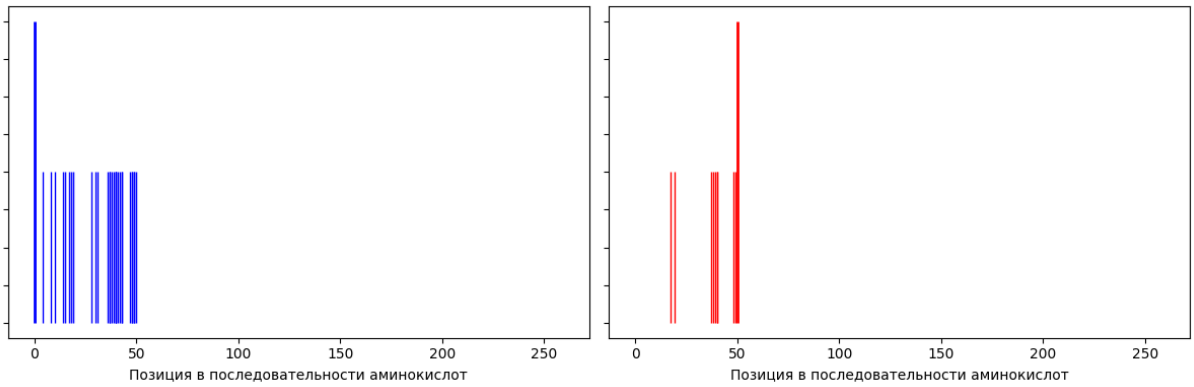

Рис. 3: Прекурсоров тут является протеолитический фрагмент, являющийся начальным фрагментам белка. Слева мы имеем *лесенки* из b-ионов, справа *лесенки* из y-ионов.