**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ**

**Кафедра биохимии**

**ОТЧЁТ ПО ПРОИЗВОДСТВЕННОЙ ПРЕДДИПЛОМНОЙ ПРАКТИКЕ**

Щур Вероники Владимировны  
студентки 5 курса 4 группы

Научный руководитель  
кандидат химических наук  
Янцевич А.В.

Минск, 2017

Содержание

Содержание 2

Введение 3

Глава 1. Обзор литературы 4

1.1 Тандемная масс-спектрометрия и фрагментация ионов-предшественников 4

1.2 Алгоритмы ПО для *de novo* секвенирования пептидов и оценочные функции 7

Глава 2. Материалы исследования 11

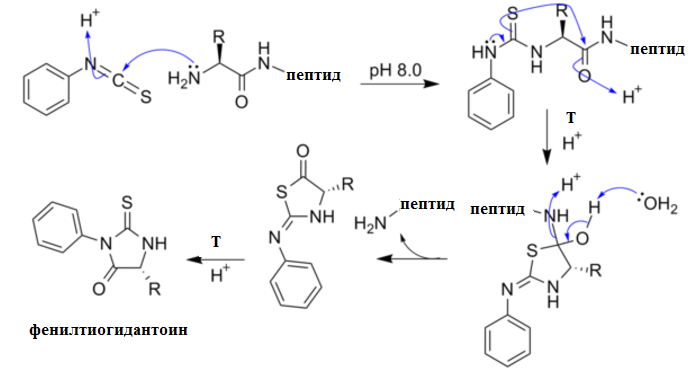
Глава 3. Результаты 12

Выводы 13

Список использованных источников 14

Введение

Определение аминокислотной последовательности пептидов, полученных после протеолиза белка, является одной из задач протеомики. Ранее для её решения использовали метод Эдмана [3]. Исследуемый пептид вступал в реакцию с фенилизотиоцианатом (ФИТЦ) [1, 3]. Этот реагент взаимодействует в слабощелочной среде с незаряженной N-концевой аминогруппой пептида с образованием производного, в котором дестабилизирована связь между альфа-карбоксильной группой N-концевой аминокислоты и альфа-аминогруппой следующей за ней аминокислоты [1, 3]. Эта связь гидролизуется без повреждения остальных пептидных связей, а комплекс ФИТЦ с N-концевой аминокислотой идентифицируют хроматографически [1]. Далее этот процесс повторяют с укороченным пептидом [1, 3]. Недостатки метода: длительность, трудоёмкость, N-конец пептида не должен быть модифицирован (ацетилирован, например), невозможно определять положение дисульфидных мостиков, концентрация пептида должна быть выше 1 пМ [1].



**Рисунок 1 – Механизм реакции, лежащей в основе метода Эдмана [1]**

На сегодняшний день для определения аминокислотной последовательности пептидов используют тандемную масс-спектрометрию (МС/МС) [6]. Существует два основных подхода: поиск по базам данных и *de novo* секвенирование [6]. Подход, основанный на использовании баз данных, имеет ограничения. Так, например, поиск осуществляется только среди уже известных белков, т.е. тех, которые есть в базе данных [4]. Мутации в генах, кодирующих белок, который есть в базе данных, не позволят его идентифицировать [4]. Аналогично в случае наличия модифицированных аминокислот и различных вариантов белка, образовавшихся после альтернативного сплайсинга мРНК [4]. Этих недостатков лишено *de novo* секвенирование пептидов.

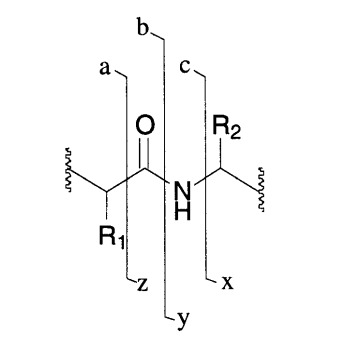
Глава 1

Обзор литературы

1.1 Тандемная масс-спектрометрия и фрагментация ионов-предшественников

Термин тандемная масс-спектрометрия означает, что используют несколько масс-анализаторов, и при переходе ионов от одного из них ко второму происходит фрагментация выбранных ионов [6]. Из смеси ионов пептидов, полученных в ходе ионизации, первый масс-анализатор отбирает обязательно многозарядные ионы с определённым m/z (это ионы-предшественники – precursor ions) [6]. Затем эти ионы подвергаются распаду на нейтральные и заряженные фрагменты в ходе CID (collision-induced dissociation – диссоциация, индуцированная столкновением) (могут быть и иные способы индукции фрагментации) [6]. Заряжённые фрагменты (product ions) анализирует второй масс-анализатор, и получается спектр этих фрагментов (МС/МС спектр) [6]. Фрагментация ионов-предшественников идёт по определённым правилам, что позволяет использовать информацию из МС/МС спектра для определения аминокислотной последовательности иона-предшественника по базам данных или без их использования (de novo секвенирование пептида) [6].

Три вида связей в составе пептида могут подвергаться фрагментации: алкилкарбонильная, пептидная и аминоалкильная [6]. В результате могут образовываться 6 типов ионов: 3 с N-конца (a, b, c) и 3 с С-конца (x, y, z) (рисунок 2) [6]. При использовании Q-TOF, тройного квадруполя или ионной ловушки в качестве масс-анализаторов преимущественно расщепляется пептидная связь, поэтому в основном образуются только 2 типа ионов: b- и у-ионы [6]. b-ионы – фрагменты, у которых положительный заряд расположен на N-конце иона-предшественника, т.к. N-концевые аминокислоты пептида протонированы при кислых рН [6]. Положительный заряд у-ионов находится на С-конце иона-предшественника, т.к. после трипсинолиза на С-конце расположены Арг или Лиз, которые также протонированы при кислых рН [6]. a-, c-, x- ,z- ионы также образуются и усложняют интерпретацию спектра [6].

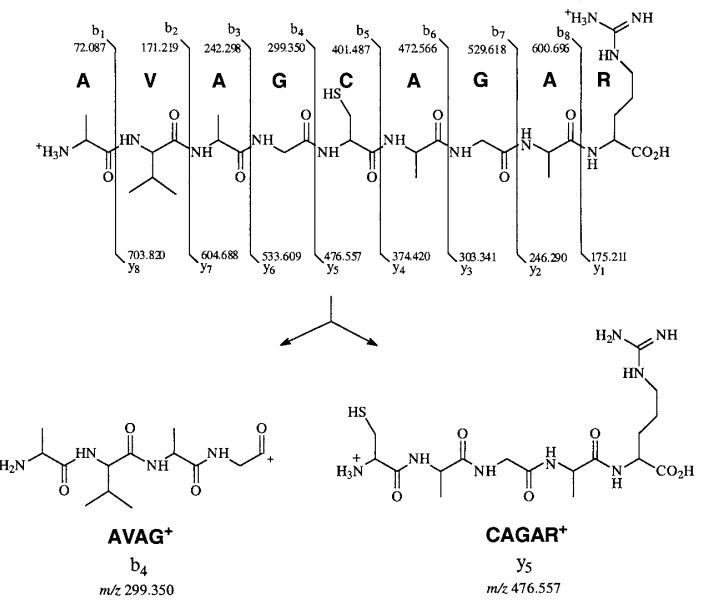


**Рисунок 2 – Номенклатура фрагментов пептидного иона-предшественника [6]**

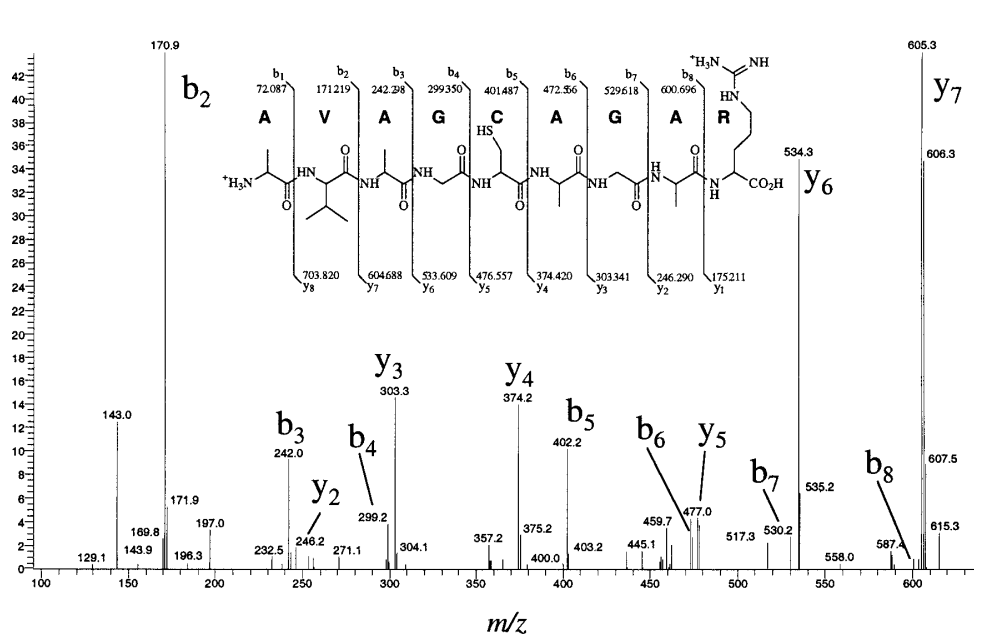
Фрагментацию пептидов при низкоэнергетичной CID объясняют концепцией мобильного протона [8]. В растворе пептид может быть протонирован по следующим сайтам: N-концевая аминогруппа пептида, аминогруппа Лиз, имидазольное кольцо Гис, гуанидиновая группа Арг [8]. Все сайты имеют приблизительно равные основности [8]. Исключение – Арг, который связывает протон наиболее прочно, что лишает последнего мобильности [8]. Таким образом, в отсутствии Арг достаточно низкоэнергетичной CID, чтобы протонирование происходило по любому из вышеназванных сайтов [8]. Мобильный протон переходит от одной пептидной связи к следующей и вызывает их расщепление, что приводит к образованию серий из b- и у-ионов [8]. Такой спектр несложно интерпретировать [8]. Если же число остатков Арг соответствует количеству протонов, то все протоны будут прочно связаны с Арг и не смогут перемещаться [8]. Такой спектр будет атипичен, пептиды будет сложно секвенировать [8]. Даже если мобильный протон есть, но в середине пептида расположен Арг, то расщепление вблизи Арг маловероятно, т.е. b- и у-ионы этого участка на спектре выглядят как пики низкой интенсивности, что также затрудняет секвенирование таких пептидов [8]. Исходя из всего вышесказанного, понятно, почему так распространено применение трипсина для получения пептидов в случае дальнейшего их анализа с использованием низкоэнергетичной CID [8]. Ведь большинство продуктов трипсинолиза белка имеют не более одного остатка Арг на С-конце. Такие пептиды двухзарядны: один протон прочно связан с боковой цепью Арг, а второй протон связан с N-концом пептида, он подвижен и способен индуцировать фрагментацию пептида [8].

На рисунке 3 и 4 представлены предполагаемая фрагментация модельного пептида AVAGCAGAR и его аннотированный МС2-спектр соответственно [6]. Модельный пептид двухзарядный: один из протонов – на аминогруппе Ала, расположенного на N-конце, а второй – на гуанидиновой группе Арг на С-конце пептида (рисунок 3) [6]. Миграция протона к пептидной связи в двухзарядном предшественнике приводит к её расщеплению, в случае модельного пептида это связь между Гли и Цис [6]. После этого формируются однозарядные b4- и y5-ионы, однако их сайты протонирования отличаются от таковых в ионе-предшественнике (рисунок 3) [6]. Но на самом деле в газовой фазе в ходе CID присутствуют и другие формы продуктов фрагментации, протонированные иначе [6].

Спектр на рисунке 4 позволяет определить аминокислотную последовательность пептида [6]. Например, разность m/z между y6 и y7 равна 71, что соответствует молекулярной массе Ала, разность m/z между у6 и у5 равна 57 – это Гли, между у5 и у4 равна 103 – это Цис и т.д. [6]. Молекулярные массы аминокислотных остатков рассчитываются, как разность молекулярной массы аминокислоты и молекулы воды, ведь при полимеризации аминокислот теряется молекула воды [6]. Серия у-ионов от у8 до у1 описывает последовательность VAGCAGAR [6]. b-серия ионов от b1 до b8 соответствует AVAGCAGA [6]. Таким образом, b- и у-серии ионов описывают одну и ту же аминокислотную последовательность, но в двух разных направлениях [6].



**Рисунок 3 – Возможная фрагментация пептида AVAGCAGAR. Структурные формулы b4- и y5-ионов – продуктов расщепления пептидной связи между G и C [6]**



**Рисунок 4 – Аннотированный МС2-спектр [M+2H]2+ иона пептида AVAGCAGAR с обозначенными b- и y-ионами [6]**

Таким образом осуществляют *de novo* секвенирование пептидов вручную, однако существует программное обеспечение позволяющее автоматизировать этот процесс [6].

1.2 Алгоритмы ПО для *de novo* секвенирования пептидов и оценочные функции

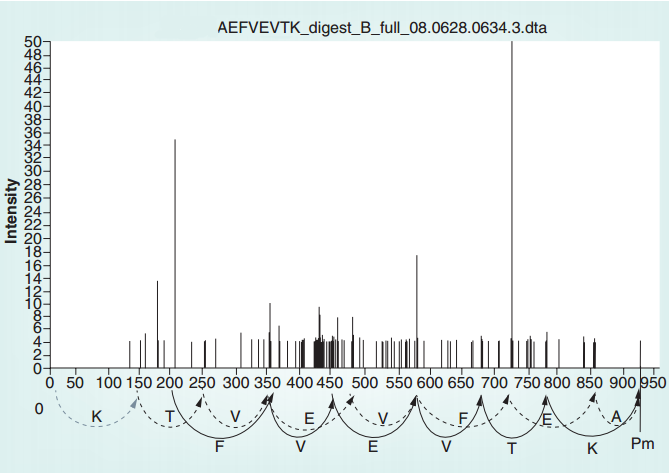
В основе функционирования ПО для *de novo* секвенирования пептидов лежат различные алгоритмы.

Первый из появившихся алгоритмов основан на создании списка последовательностей пептидов-кандидатов, имеющих такую же массу, как и ион-предшественник из экспериментального спектра [7, 8]. Затем происходит сравнение предсказанных фрагментов пептидов-кандидатов с фрагментами из экспериментального спектра [7, 8]. Недостаток такого алгоритма – генерация огромного количества пептидов-кандидатов, поэтому подход применим лишь для секвенирования пептидов не длиннее 8 аминокислотных остатков [8].

Второй вариант алгоритма основан на сравнении с экспериментальным спектром коротких аминокислотных последовательностей, представляющих собой лишь участок искомого пептида [7, 8]. Затем такая короткая последовательность удлиняется на 1 аминокислотный остаток, после чего снова происходит сравнение её с экспериментальным МС/МС-спектром [7, 8]. В течение такого постепенного удлинения сохраняются лишь те последовательности, чей предсказанный спектр фрагментации наиболее близок к экспериментальному [7, 8]. В этом и кроется недостаток алгоритма: возможно ошибочное исключение «удачных» пептидов-кандидатов из-за того, что в экспериментальном спектре некоторые участки пептида могут быть представлены слабоинтенсивными пиками или вообще не представлены [7].

Третий подход напоминает ручное секвенирование, но с использованием графического дисплея [7, 8]. Дисплей показывает, как различные заряженные фрагменты пептида соединены с другими. Фрагменты могут быть соединены, только если разность их масс соответствует массе какого-либо аминокислотного остатка [7, 8]. Пользователь может выбрать вариант соединения фрагментов, при котором они будут отличаться на массу аминокислотного остатка и суммарная масса пептида будет соответствовать массе иона-предшественника [8].

Четвёртый тип алгоритмов наиболее распространённый и основан на теории графов [7, 8]. Алгоритм использованного нами ПО относится к этому типу. Программа осуществляет трансформацию каждого пика экспериментального спектра в вершину (узел) графа спектра [2, 7, 8]. Две вершины могут быть соединены рёбром, только если разность m/z между ними соответствует массе 1 или нескольких аминокислотных остатков [2, 7, 8]. Далее необходимо найти путь через соединённые вершины, чтобы определить аминокислотную последовательность [2, 7, 8]. На рисунке 5 – граф спектра с сериями b- и у-ионов [2]. Обе серии непрерывны – это признак верного предсказания последовательности [2].



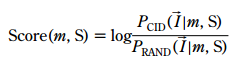
**Рисунок 5 – Аннотированный МС/МС-спектр пептида AEFVEVTK. Однозарядные b- и у-серии ионов (сплошные и прерывистые линии соответственно) соединены рёбрами, под которыми указаны соответствующие аминокислоты. Для ясности не представлены другие варианты соединения пар ионов. Pm – масса иона-предшественника [2]**

Цель интерпретации МС/МС спектра – обнаружить пептид, чей гипотетический спектр фрагментации наиболее схож с экспериментальным [7]. Чтобы оценить степень сходства используют оценочную функцию (score), дизайн которой отличается в разном ПО, даже если алгоритмы программ относятся к одному и тому же типу (например, используют теорию графов) [7].

В PepNovo+ – ПО, которое мы использовали, – есть 2 оценочные функции: PnvScr (PepNovo score) и RnkScr (ranking score).

Основа расчёта PnvScr – сравнение 2 гипотез, относящихся к спектру S и массе m возможного сайта расщепления [4]. CID-гипотеза: m – подлинное расщепление в пептиде, которому соответствует спектр S [4]. Существуют правила фрагментации пептидов, поэтому есть определённые более вероятные комбинации масс фрагментов и их интенсивностей [4]. Разработчики ПО смоделировали эти правила фрагментации для того, чтобы определить вероятность детекции наблюдаемого набора интенсивностей фрагментов , а m – расщепление в пептиде, которому соответствует спектр S [4]. Гипотеза случайных пиков (random peaks hypothesis – RAND): пики с указанными интенсивностями появились в спектре благодаря случайным событиям (т.е. нет никаких правил, объясняющих их появление) [4].

PnvScr – десятичный логарифм отношения вероятностей этих 2 гипотез:



Большинство оценочных функций (в том числе и выше приведенная) создают вероятностную модель, которая хорошо работает лишь в случае малого пространства поиска (поиск среди небольшого набора последовательностей) и не отличает так называемые гомеометрические пептиды (состоят из разных аминокислотных последовательностей, но имеют сходные гипотетические спектры фрагментации) [5]. Поэтому разработчики PepNovo+ ввели ещё одну оценочную функцию (RnkScr), которая решает эти проблемы [5]. Для её создания использовали машинное обучение на большом наборе данных [5].

Глава 2

Материалы исследования

1. МС/МС спектры следующих белков:

* БСА – бычий сывороточный альбумин,
* ЧСА – человеческий сывороточный альбумин,
* Lmw b5-Lmw CPR и Lmw b5-Hmw CPR – рекомбинантные сшитые белки, состоящие из цитохрома b5 (cyt b5) и НАДФН-цитохром Р450 редуктазы (CPR). Lmw (low molecular weight) и Hmw (high molecular weight) отличаются отсутствием и наличием гидрофобного домена соответственно.

Остатки цистеина в белках были восстановлены (обработка 1,4-дитиотреитолом) и карбамидометилированы (обработка 2-иодацетамидом). Белки подвергли трипсинолизу в растворе. Перед ВЭЖХ-МС/МС проводилась твёрдофазная экстракция. Спектры были получены ранее в ходе выполнения курсовой работы.

2. Компьютер с установленным ПО:

* Agilent MassHunter Workstation. Qualitative Analysis – получение МС/МС спектров в формате mzdata.xml;
* Spectrum Mill – поиск белков, соответствующих полученным пептидам, в базе данных Swiss-Prot;
* mz2mgf – конвертация mzdata.xml в mgf;
* PepNovo+ – программа для *de novo* секвенирования пептидов.

Глава 3

Результаты

Результаты представлены в таблице 1. Таблица состоит из 3 основных столбцов, каждый из которых подразделяется ещё на несколько. Пояснения приводим ниже.

1. Секвенирование по SwissProt

* Длина последовательности – число аминокислотных остатков в последовательности пептида.
* Последовательность – аминокислотная последовательность пептида, полученная с помощью Spectrum Mill.
* Spectrum Mill Score – значение оценочной функции, характеризующее сходство экспериментального спектра и спектра из базы данных SwissProt.
* Средний Spectrum Mill Score ± SD – среднее арифметическое Spectrum Mill Score ± его стандартное отклонение. SD не рассчитано адекватно для последовательностей длиной в 14, 15 и 21 аминокислотный остаток, т.к. они представлены в единственном повторе.

2. *De novo* секвенирование

* Лучшая предсказанная последовательность – последовательность, предсказанная PepNovo+ и наиболее похожая на ту, которая получена с помощью Spectrum Mill. Полужирным выделены верные аминокислотные остатки (см. ниже).
* RnkScr, PnvScr – значения оценочных функций в PepNovo+.

3. Сравнение методов секвенирования

* N(верных АК) – количество аминокислотных остатков, которые совпали при сравнении последовательностей, полученных в ходе поиска по базе данных и *de novo* секвенирования.
* Точность предсказания, % = [N(верных АК)/Длина последовательности]\*100%
* Средняя ТП ± SD, %– среднее арифметическое точности предсказаний последовательностей одной длины ± его стандартное отклонение. SD не рассчитано адекватно для последовательностей длиной в 14, 15 и 21 аминокислотный остаток, т.к. они представлены в единственном повторе.

Выводы

Анализируя данные таблицы 1 можно сделать следующие выводы.

1. С увеличением длины последовательности снижается точность предсказания с помощью PepNovo+. Максимальная средняя точность предсказания наблюдается для пептидов длиной в 10 (85,00±21,21%) и 8 аминокислотных остатков (84,38±18,75%). Однако высокие значения стандартных отклонений не позволяют говорить уверенно о найденной оптимальной для PepNovo+ длине пептида.

2. С увеличением длины пептида PepNovo+ хуже предсказывает N-концевые, а с последующим увеличением N- и С-концевые аминокислотные остатки. Возможная причина – качество МС/МС спектра.

3. Точность предсказания PepNovo+ была бы выше, если бы не такие ошибки *de novo* секвенирования, как неверное положение 2 соседних аминокислотных остатков (VK**PQVSTPTLVE**DVK, а нужно KVPQVSTPTLVEVSR), а также замена верной аминокислоты на неверную, но с такой же массой (например, Иле на Лей в случае L**QTTAPPVK**, **AP**L**LSDSS** а нужно IQTTAPPVK, APILSDSSCK).

4. Не наблюдается взаимосвязи между Spectrum Mill Score и точностью предсказаний PepNovo.

5. Не наблюдается взаимосвязи между точностью предсказания и оценочными функциями PepNovo+.

список использованных источников

1. *Овчинников, Ю.А.* Биоорганическая химия / Ю. А. Овчинников. М.: Просвещение, 1987. – 816 с.

2. *Allmer, J.* Algorythms for de novo sequencing of peptides from tandem mass spectra // Expert Rev Proteomics. – 2011. – Vol. 8(5). – P. 645–657.

3. *Edman, P., Begg, G.* A protein sequenator // Eur. J. Biochem. – 1967. – Vol. 1(1). – P. 80–91.

4. *Frank, A., Pevzner P.* De Novo Peptide Sequencing via Probalistic Network Modeling // Anal.Chem. – 2005. – Vol. 77. – P. 964–973.

5. *Frank, A.M.* A Ranking-Based Scoring Function For Peptide-Spectrum Matches // J. Proteome Res. – 2009. – Vol. 8(5). – P. 2241–2252.

6. *Liebler, D.C.* Introduction to Proteomics. Tools for the New Biology / D.C. Liebler. Totowa. NJ: Humana Press, 2002. – 198 p.

7. *Lu, B., Chen, T.* Algorythms for de novo peptide sequencing using tandem mass spectrometry // DDT: BIOSILICO. – 2004. – Vol.2, No.2. – P. 85–90.

8. *Ma, B., Johnson, J.* De Novo Sequencing and Homology Searching // Molecular & Cellular Proteomics. – 2012. – Vol. 11.2.