

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ
УЧРЕЖДЕНИЕ «НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ
БИОМЕДИЦИНСКОЙ ХИМИИ ИМЕНИ В.Н. ОРЕХОВИЧА»**

На правах рукописи

ИЛЬГИСОНИС ЕКАТЕРИНА ВИКТОРОВНА

**ПРОТЕОТИПИЧЕСКИЕ ПЕПТИДЫ ДЛЯ
КОЛИЧЕСТВЕННОГО МАСС-СПЕКТРОМЕТРИЧЕСКОГО АНАЛИЗА БЕЛКОВ,
КОДИРУЕМЫХ ГЕНАМИ ХРОМОСОМЫ 18 ЧЕЛОВЕКА**

03.01.09 – математическая биология, биоинформатика

**Диссертация на соискание ученой степени
кандидата биологических наук**

**Научный руководитель:
д. б. н., член-корр. РАН Лисица А.В.**

Москва 2015

Оглавление

1. ВВЕДЕНИЕ, ЦЕЛЬ И ЗАДАЧИ	4
2. ОБЗОР ЛИТЕРАТУРНЫХ ИСТОЧНИКОВ	7
2.1. Масс-спектрометрия в протеомике	7
2.2. Схема масс-спектрометрического эксперимента	8
2.3. Метод масс-спектрометрического мониторинга диссоциативных переходов	10
2.4. Планирование направленного масс-спектрометрического эксперимента	13
2.5. Интерференция в масс-спектрометрии	16
2.6. Количественное измерение содержания белков	18
2.7. Стандартизация результатов протеомных исследований	20
2.8. Программные системы для разработки и анализа результатов масс-спектрометрических экспериментов	24
2.9. Репозитории результатов масс-спектрометрических экспериментов	26
2.10. Использование метода мониторинга диссоциативных переходов в экспериментально-клинических исследованиях	29
2.11. Хромосомотцентричный подход	32
3. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ	33
3.1. Исходные масс-спектрометрические данные	33
3.2. Первичная обработка данных, полученных методом мониторинга диссоциативных переходов	34
3.3. Проверка формата входных данных	37
3.4. Программная реализация базы данных протеотипических пептидов	38
3.5. Загрузка результатов масс-спектрометрических экспериментов в базу данных	40
3.6. Формирование запросов к базе данных	42
3.7. Оценка результатов масс-спектрометрических измерений	44
3.8. Используемые информационные ресурсы	47
3.9. Специализированное программное обеспечение	49
4. ОСНОВНЫЕ РЕЗУЛЬТАТЫ	52
4.1. Структура данных для описания результатов направленного масс-спектрометрического эксперимента	52
4.2. Система запросов к информации о протеотипических пептидах	60
4.3. Оценка результатов детекции белков и пептидов в биологических образцах	70
4.4. Анализ выборки протеотипических пептидов и соответствующих им белков	77
4.4.1. Репрезентативность выборки протеотипических пептидов и соответствующих им белков	77
4.4.2. Интерференция протеотипических пептидов	85
5. ЗАКЛЮЧЕНИЕ	99
6. ВЫВОДЫ	103
7. СПИСОК СОКРАЩЕНИЙ	104

8. ОСНОВНЫЕ ПОНЯТИЯ И ТЕРМИНЫ	105
9. СПИСОК ЛИТЕРАТУРЫ.....	107
ПРИЛОЖЕНИЕ	118

1. ВВЕДЕНИЕ, ЦЕЛЬ И ЗАДАЧИ

Повышение эффективности методов анализа биологических молекул за последние 20 лет сделало возможным выполнение широкомасштабных проектов, таких как «Геном человека» [Venter и др., 2001], «Протеом человека» [Orchard, Ping, 2009]. Ключевая характеристика широкомасштабных научных проектов состоит в том, что в результате их выполнения не появляется исчерпывающая информация об исследуемом объекте, а формируется вектор будущих узконаправленных исследований в виде массива накопленных сведений [Pearsons, 1991]. Результаты широкомасштабных проектов требуют интерпретации и уточнения.

Особенность постгеномных проектов позволяет рассматривать не совокупность экспериментов, направленных на подтверждение научной гипотезы, а как исследовательский процесс. Исследовательский процесс – это упорядоченная последовательность операций, выполнение которых приводит к определенному результату. Для построения структуры данных исследовательского процесса необходимо определить критерии качества его реализации.

В 2010 был дан старт международному хромосомотцентричному проекту «Протеом человека», в котором научные консорциумы из разных стран исследуют белки, кодируемые генами одной из выбранных соответствующей страной хромосом человека [Paik и др., 2012]. С точки зрения диапазона концентраций белков и их участия в молекулярных процессах, все хромосомы примерно равнозначны [Ponomarenko и др., 2012], поэтому методы исследований, отработанные на одной хромосоме, переносимы и на другие хромосомы.

Каждая хромосома содержит белки, линейный диапазон концентраций которых составляет от 10^{-18} до 10^{-6} копий на клетку [Archakov и др., 2011]. Для достижения сверхнизких концентраций основным инструментом проекта «Протеом человека» был выбран направленный масс-спектрометрический подход [Archakov и др., 2011].

Россия как страна-участница международного проекта «Протеом человека», завершила масс-спектрометрические измерения для белковых продуктов

хромосомы 18 человека [Ponomarenko и др., 2014; Пономаренко и др., 2015]. Несмотря на высокую чувствительность метода мониторинга диссоциативных переходов, достоверность детекции падает из-за возникновения интерференции [Bao и др., 2013; Sherman и др., 2009; Sherman, Molloy, Burlingame, 2012].

Интерференцией обозначают явление, когда регистрируемый масс-спектрометрический сигнал соответствует не одному пептиду, а нескольким неизвестным компонентам сложной биологической матрицы. До сих пор в практике направленной масс-спектрометрии отсутствуют средства статистической оценки правдоподобия детекции пептида, и решения выносятся на основе экспертного мнения. Оценка достоверности результатов мониторинга диссоциативных переходов предусматривает планирование эксперимента в соответствии с требованиями конкретного ПО. Разработка алгоритмов для этих программ проводилась на обучающих выборках, состоящих только из высококопийных белков-биомаркеров [MacLean и др., 2010; Reiter и др., 2011].

Накопленный в ходе выполнения российской части проекта «Протеом человека» массив данных является уникальным, поскольку он получен с использованием хромосомотцентричного подхода. Это означает, что объектами исследования являются белки, которые кодируются генами, локализованными на одной хромосоме. В данной работе впервые проводится анализ столь объемного, но в то же время, гомогенного массива экспериментальных данных, полученных на одном приборе по стандартному протоколу в течение полутора лет.

Ранее предложенные алгоритмы оценки достоверности результатов экспериментов, полученных методом мониторинга диссоциативных переходов [Reiter и др., 2011], основывались на обучающих выборках и требовали внесения изменений в протокол эксперимента. Используемый в данной работе алгоритм фильтрации и оценки воспроизводимости экспериментальных результатов не зависит от обучающей выборки и протокола эксперимента.

Обработка однородного массива экспериментов с целью выявления закономерностей детекции белков может послужить основой для стандартизации протокола анализа результатов мониторинга диссоциативных переходов.

Настоящая работа направлена на формализацию критериев оценки результатов детекции методом мониторинга диссоциативных переходов для создания базы данных протеотипических пептидов для количественного анализа белков. В работе показано, что хромосомоцентричный подход позволяет получить универсальные критерии, применимые к оценке качества масс-спектрометрических измерений количественного содержания любых белков. Полученная база данных протеотипических пептидов позволит осуществлять аннотацию биологического материала с точки зрения содержания в них протеотипических пептидов, результаты детектирования которых обладают наибольшей воспроизводимостью.

Цель работы: Разработать алгоритм анализа результатов масс-спектрометрических измерений протеотипических пептидов белков человека для формирования базы данных протеотипических пептидов на примере хромосомы 18.

Задачи:

1. Провести анализ предметной области направленной масс-спектрометрии; выявить информационные объекты и взаимосвязи, на их основе разработать структуру данных.
2. Разработать алгоритмы автоматизированной обработки экспериментальных данных с определением уровня воспроизводимости детекции пептидов методом МДП.
3. Провести оценку результатов масс-спектрометрических измерений протеолитических пептидов белков хромосомы 18 и сформировать контрольную выборку протеотипических пептидов для количественных измерений соответствующих белков в биологических пробах.
4. Провести обработку экспериментальных данных масс-спектрометрического анализа проб плазмы крови человека и клеточной линии HepG2 с использованием биоинформатических ресурсов.

2. ОБЗОР ЛИТЕРАТУРНЫХ ИСТОЧНИКОВ

Задача протеомики состоит в идентификации и количественном измерении содержания белков в биологических образцах сложного состава [Patterson, Aebersold, 2003]. Наиболее чувствительным и производительным методом решения этой задачи является хромато-масс-спектрометрия [Aebersold, Mann, 2003].

Перспективной стратегией протеомики является сочетание хромосомоцентричного подхода с методом направленной масс-спектрометрии [Paik и др., 2012]. В отличие от общепринятого подхода, при котором эксперименту предшествует формулирование научной гипотезы и определение целевой группы исследуемых белков, хромосомоцентричный подход предполагает последовательную инвентаризацию всех белков протеома методом мониторинга диссоциативных переходов (МДП). Размер, то есть количество исследуемых объектов – белков, протеома подразумевает формирование инструкций по выполнению этапов исследования, а также разработку системы стандартов выполняемых операций. Для интерпретации результатов экспериментов должны быть сформулированы критерии, в соответствии с которыми будет проводиться оценка экспериментальных результатов.

В обзоре изложены существующие на сегодняшний день стандарты проведения протеомных исследований масс-спектрометрическим методом МДП, а также проблемы стандартизации научных исследований на примере масс-спектрометрического анализа протеома человека.

2.1. *Масс-спектрометрия в протеомике*

Предметом измерения в масс-спектрометрии является отношение массы к заряду исследуемых ионов (m/z , *масс-зарядная характеристика*) [Aebersold, Mann, 2003]. Результатом масс-спектрометрического измерения является масс-спектр, который представляет собой распределение заряженных частиц по их массово-зарядовым числам.

Преимуществом масс-спектрометрии по сравнению с другими аналитическими методами является возможность получения информации о сотнях биомолекул одновременно. Если оптические, рентгеновские методы основаны на детекции излучения или поглощения энергии молекулами или атомами, то в основе масс-спектрометрии лежит детекция свойств вещества, его массы и заряда. По сравнению с иммунохимическими методами детекции белков в масс-спектрометрическом методе мониторинга диссоциативных переходов возможность перекрестных реакций с другими белками существенно ниже [Addona и др., 2009], тогда как результаты измерений совпадают на 90-99% [Liu и др., 2012].

Масс-спектрометрические методы рутинно применяются для анализа низкомолекулярных биологических молекул, в том числе лекарств, наркотиков, допинговых препаратов, метаболитов [Zweigenbaum, Henion, 2000]. К настоящему моменту получено более 120 000 масс-спектров для нескольких тысяч органических и неорганических соединений¹ (включая метаболиты, аминокислоты, сахара, липиды, лекарственные средства, пестициды и т.п.).

2.2. Схема масс-спектрометрического эксперимента

Одним из широко используемых подходов в масс-спектрометрических протеомных исследованиях является пептидоцентричный подход, или подход «снизу-вверх», при котором информацию о белках получают, основываясь на детекции пептидов [Duncan, Aebersold, Caprioli, 2010].

В качестве аналита используется сложная смесь пептидов, полученная в результате ферментативного расщепления белковой фракции биологического образца. Наиболее часто применяемым для протеолиза ферментом является трипсин. Трипсин специфичным образом гидролизует пептидную связь с карбоксильного конца только в двух положениях: после аргинина (R) или лизина (K) [Keil, 1992; Olsen, Ong, Mann, 2004; Rodriguez и др., 2007]. Основные аминокислоты на концах триптических пептидов способствуют образованию

¹ URL: <http://www.nist.gov/mml/bmd/data/tandemmass-specplib.cfm>

положительно заряженных пептидных ионов. Смесь продуктов трипсинолиза разделяется с помощью системы высокоэффективной жидкостной хроматографии (ВЭЖХ), которая в режиме реального времени сопряжена с масс-спектрометром. Каждая фракция, образовавшаяся после разделения, подвергается масс-спектрометрическому анализу.

Масс-спектрометрический анализ проводится в приборе, который состоит из источника ионов, масс-анализатора, ячейки соударений, и, наконец, детектора. Реализация и физический механизм действия этих составляющих может меняться в зависимости от типа прибора и производителя, однако общая схема эксперимента сохраняется. Сортировка ионов по массово-зарядовым числам происходит в масс-анализаторе прибора. Масс-анализатор представляет собой камеру, в которой поддерживается вакуум и создается напряженность электрического поля. Под действием электрического поля ионы начинают двигаться по направлению от подсоединенного к ВЭЖХ электроспрейного капилляра ионизатора к детектору. По пути движения ионов в камере электронные поля выстроены таким образом, чтобы задерживать или наоборот пропускать к детектору ионы с определенным соотношением массово-зарядовых характеристик [Yost, Enke, 1978].

В протеомике используют два хромато-масс-спектрометрических подхода, различающихся по принципу масс-спектрометрической детекции: панорамную и направленную масс-спектрометрию.

Панорамная (в англоязычных источниках – shotgun) масс-спектрометрия используется для анализа заранее неизвестных компонентов биологического образца. Этот режим еще называют информационно-зависимым (в англоязычной литературе – data-dependent mode). Информационно-зависимый режим не требует этапа биоинформатической подготовки к эксперименту, обязательным условием является биоинформатическая интерпретация полученных результатов [Govorun, Archakov, 2002]. Идентификация зарегистрированных белков осуществляется посредством сопоставления полученных спектров пептидов с базой данных и расчета вероятности

ложноположительных identifications [Koenig и др., 2008; Perkins и др., 1999]. Эксперименты в панорамной протеомике направлены на идентификацию максимального числа белков в широком динамическом диапазоне концентраций [Kim и др., 2014; Wilhelm и др., 2014a].

В информационно-зависимом режиме масс-спектрометр регистрирует сигнал от всех пептидов, присутствующих в пробе. Направленные масс-спектрометрические эксперименты (в англоязычной литературе - data-independent mode), наоборот, предназначены для детекции заранее заданных исследователем масс-зарядных характеристик. Метод мониторинга диссоциативных переходов основан на масс-спектрометрической регистрации заряженных ионов, возникающих в результате ионизации и фрагментации пептидов, соответствующих заданному исследователем набору белков [Domon, 2012; Picotti, Aebersold, 2012; Zhi, Wang, She, 2011]. Этот режим приводит к повышению чувствительности метода на два порядка по сравнению с панорамными, позволяет детектировать белки в концентрационном диапазоне до пяти порядков [Lange и др., 2008a]. Показано, что метод МДП позволяет анализировать низкокопийные белки в аттомолярных концентрациях в сложных биологических образцах [Korylov и др., 2013].

2.3. *Метод масс-спектрометрического мониторинга диссоциативных переходов*

Отличием метода мониторинга диссоциативных переходов от панорамных методов является необходимость биоинформатической подготовки. Перед проведением эксперимента производится расчет масс-зарядных характеристик выбранного пептидного иона и его фрагментов.

В соответствии с расчетом производится настройка масс-анализаторов, квадрупольных фильтров. В результате первичной фильтрации масс-зарядных характеристик пептидных ионов в ячейку соударений проходят только ионы, масс-зарядные соотношения которых удовлетворяют заданному диапазону m/z . Прошедшие первый барьер ионы аккумулируются во втором квадруполе (ячейка соударений) в количестве, достаточном для получения сигнала на детекторе. Накапливаемые ионы подвергаются диссоциации, в результате чего

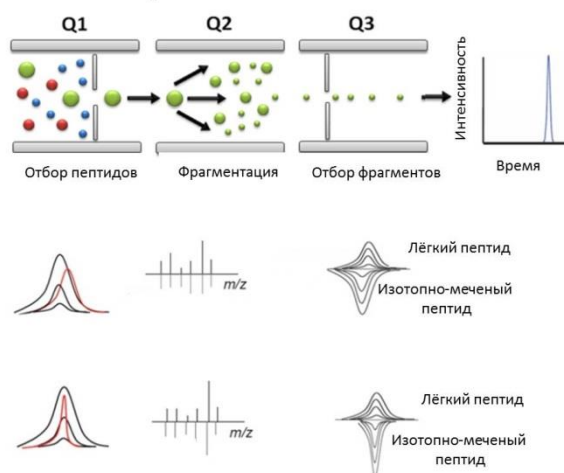
образуются фрагменты пептидного иона. Для диссоциации используются молекулы инертного газа, которые, сталкиваясь с пептидным ионом, передают ему дополнительную энергию. Эта энергия приводит к разрыву пептидных связей по множеству возможных сценариев фрагментации (минимум, 30 – по данным Spectrum Mill²). Из полученных фрагментов фильтр следующего квадрупольного отсеивает на пути к детектору ионы, не удовлетворяющие заданному отношению массы к заряду. Таким образом, детектируется сигнал, соответствующий только заданным исследователем пептидам.

²URL:http://proteomics.broadinstitute.org/millbin/msprod.cgi?msparams_dir=msparams_mill%2F&parent_mass_convert=monoisotopic&max_ms_prod_charge=2&sequence=AAAAAAAAAK&it=b&it=y&it=I&user_aa_c=2&user_aa_h=3&user_aa_n=1&user_aa_o=1&user_aa_s=0&user_aa_p=0

Планирование эксперимента



Экспериментальная часть



Интерпретация и публикация данных



Рисунок 1. Схема протеомного эксперимента, осуществляемого методами направленной масс-спектрометрии. Фрагменты схемы адаптированы из статей [Schmidt, Picotti, Aebersold, 2008; Surinova и др., 2013].

Q1-Q3 квадрупольные фильтры масс-анализатора; PASSEL PANORAMA – репозитории для хранения результатов измерений, полученных методом МДП.

Масс-спектрометрический сигнал, получаемый от пептидного иона, представляет собой группу хроматограмм коэлюирующих соединений, соответствующих фрагментам пептидного иона. Поскольку исходный пептид распадается на фрагменты в ячейке соударений тройного квадрупольного поля, на хроматограмме отсутствует соответствующий ему пик. Исходный пептидный

ион диссоциирует на несколько фрагментов, поэтому хроматограмма выглядит как группа коэлюирующих соединений (рис. 1).

Реконструированная хроматограмма, построенная на основе сигналов одного пептидного иона, называется экстрагированной ионной хроматограммой (в англоязычной литературе – XIC или EIC, extracted-ion chromatogram). Небольшие несовпадения между пиками (по форме или времени, которому соответствует вершина) неизбежно присутствуют в хроматограммах. Причиной возникновения таких несовпадений могут стать различные факторы, в частности время накапливания пептидных ионов во втором квадруполе (в англоязычной литературе – dwell time) или задержки, связанные с переключением режимов работы квадруполей, составляющие микросекунды.

Количество пептидов, по которым проводится масс-спектрометрическая идентификация белка в панорамных экспериментах, зависит от его концентрации: обычно, чем ниже концентрация белка, тем меньше пептидов [Liu и др., 2004]. Количество пептидов, по которым производят идентификацию белка, составляет не менее двух и может достигать до нескольких десятков в панорамных экспериментах [Nesvizhskii, Aebersold, 2005; Pirmoradian и др., 2013]. В направленных масс-спектрометрических экспериментах используют всего несколько протеотипических пептидов, которые служат «прототипами» целого белка после его ферментативного гидролиза. Эти пептиды называют протеотипическими.

2.4. *Планирование направленного масс-спектрометрического эксперимента*

Понятие протеотипического пептида является ключевым в современной биологической масс-спектрометрии [Alves, Ogurtsov, Yu, 2011]. Согласно работам [Craig, Cortens, Beavis, 2005; Mallick и др., 2007], протеотипическим считается пептид, который детектируется масс-спектрометрическими методами как минимум в половине экспериментов, в которой детектируется исходный белок. При этом в работах [Mallick и др., 2007] проводилось исследование почти 500 физико-химических параметров пептидов (в т.ч. заряд, гидрофобность) для выявления свойств пептидов, которые можно

детектировать масс-спектрометрическим способом, однако закономерностей выявить не удалось. В работе [Kuster и др., 2005] дополнительной характеристикой протеотипического пептида является его уникальность. Последовательность пептида должна быть уникальна в масштабах генома, пептид должен встречаться только в одном белке или в одной изоформе. По теоретическим расчетам, проведенным Чернобровкиным А.Л., таких пептидов, возникающих в результате трипсинолиза, в протеоме человека достаточно много – в среднем от 10 до 15 для белка массой 40 кДа. Для участка цепи длиной от семи аминокислот вероятность быть уникальным составляет около 95% [Чернобровкин, 2012]. Авторы [Selevsek и др., 2011] используют слово «протеотипический» как синоним уникального, не вкладывая в этот термин ионизационные свойства пептида. В работе [Ludwig и др., 2012] протеотипическим пептидом также считали уникальный пептид – продукт полностью прошедшего ферментативного расщепления трипсином.

Поскольку аминокислотная последовательность пептида подбирается так, чтобы она встречалась только в одном среди всех кодируемых геномом белков, то наличие на хроматограмме сигнала от этого пептида подтверждает наличие белка. Из общих соображений следует, что уникальные пептиды, соответствующие одному белку, должны находиться в эквимоллярных концентрациях [Bondarenko, Chelius, Shaler, 2002]. Однако, интенсивность масс-спектрометрического сигнала зависит не только от концентрации белка. В реальности пептиды одного и того же белка могут давать как высокий, так и низкий сигнал, что объясняется разницей в их физико-химических свойствах [Blonder, Veenstra, 2007]. Это было продемонстрировано Mallick и соавторами [Mallick и др., 2007] на примере 18 белков, которые были смешаны в различных концентрациях от 4 до 1000 наномоль. Серия из 22 наблюдения показала, что некоторые пептиды детектировались чаще, чем другие. Связи между частотой детекции пептида и концентрацией исходного белка установлено не было. Таким образом, важное свойство протеотипического пептида заключается в том, что его концентрация, установленная посредством построения

калибровочного уравнения, должна коррелировать с концентрацией исходного белка. Такую корреляцию можно установить путем сопоставления параллельных измерений масс-спектрометрическими и иммуноаффинными методами [Zhi, Wang, She, 2011].

Выбор протеотипических пептидов основывается на сведениях, содержащихся в научной литературе или библиотеках предварительно проведенных панорамных экспериментах. Как правило, используют пептиды с длиной от 7 до 25 а.о., поскольку вероятность быть уникальным у пептида, состоящего менее чем из семи а.о., мала, а масса пептида, состоящего из более чем 25 а.о., превышает массу, на которую можно настроить квадрупольный масс-анализатор. Масса двухзарядных ионов такого пептида будет превышать m/z 1300 [Crasto и др., 2013]. В работе [Alves, Ogurtsov, Yu, 2011] предлагается следующий алгоритм выбора протеотипических пептидов для направленного масс-спектрометрического эксперимента. Он включает отбор целевых белков, вовлеченных в биологический процесс (развитие заболевания, метаболический путь), получение их аминокислотной последовательности в формате FASTA [Pearson, Lipman, 1988], проведение гидролиза *in silico*, отбор пептидов нужной длины, поиск пептидных последовательностей алгоритмом BLAST [Altschul и др., 1997; Camacho и др., 2009]. При этом для количественного измерения содержания пептида предлагается использовать калибровочное уравнение.

Наиболее популярным ресурсом, предоставляющим доступ к поиску белков по последовательностям пептидов, является UniProt (Universal Protein Resource, www.uniprot.org). Ресурс содержит аннотированные аминокислотные последовательности белков человека. Источником последовательностей в базе данных UniProtKB служат белковые базы данных (PDB, Ensembl и др.), а также материалы научных публикаций. Все последовательности проходят экспертную проверку. Ресурс содержит сведения о так называемых канонических последовательностях — наиболее часто встречающихся, а также перечень сплайс-опосредованных вариантов белковых последовательностей.

2.5. *Интерференция в масс-спектрометрии*

В силу сходства своих физико-химических свойств пептиды дают сложную, интерферирующую масс-спектрометрическую картину. Причина возникновения интерференции заключается в том, что масс-зарядные характеристики пептидных ионов и их фрагментов не являются уникальными. В работе [Sherman и др., 2009] было показано, что для каждого белка, входящего в состав протеома человека, возможно отобрать пептид и его фрагменты, сочетание масс-зарядных характеристик которых будет уникальным в рамках протеома. Однако биологический образец представляет собой сложную органическую смесь. Поведение в масс-спектрометре компонентов биологического образца и модифицированных белков недостаточно изучено и потому непредсказуемо.

В отличие от панорамных экспериментов, предназначенных для обзорного анализа множества белков, в направленных протеомных экспериментах количество анализируемых пептидов невелико. Измеряется до 100 пептидов одновременно [Lange и др., 2008b], поэтому отбор целевых протеотипических пептидов и их фрагментов представляет особую важность в планировании эксперимента. Ряд подходов разработан, чтобы помочь исследователю выбрать протеотипические пептиды и соответствующие им фрагменты, а также оценить достоверность идентификации.

К примеру, в основе программы SRMCollider [Röst, Malmström, Aebersold, 2012], служащей для оптимизации протокола работы, лежит симуляция эксперимента на масс-спектрометре с тройным квадруполом. Программа ищет возможные пересечения между масс-зарядными характеристиками предварительно отобранных протеотипических пептидов целевых белков и фоновых белков биологического образца. SRMCollider позволяет приблизительно оценить хроматографическое время удерживания пептидов и указывает исследователю на теоретическую возможность возникновения интерференции, затрудняющей интерпретацию результатов анализа.

Дополнительная априорная информация о возможности детекции протеотипического пептида может быть также получена с использованием программного обеспечения, разработанного Mallick и соавторами [Mallick и др., 2007] на обучающей выборке из 46 млн пептидов, идентифицированных на различных масс-спектрометрических платформах. Предлагаемые алгоритмы учитывают особенности аминокислотного состава и физико-химические свойства потенциальных протеотипических пептидов. Несмотря на различные подходы, не удалось создать универсальный инструмент, обеспечивающий ранжирование потенциально детектируемых пептидов [Рыбина и др., 2014; Скворцов и др., 2015]. Условия ионизации, фрагментации и детекции каждого эксперимента, а также биологический материал могут оказывать критическое влияние на результаты эксперимента.

Количество пептидов, имеющих совпадающие масс-зарядные характеристики, зависит от точности их измерения прибором. Принцип метода мониторинга диссоциативных переходов заключается в использовании масс-анализатора для фильтрации ионов по масс-зарядным характеристикам. Точность измерения масс-зарядных характеристик обеспечивается шириной измерительного окна масс-анализатора. Чем уже это окно, тем выше точность измерений; минимальный размер окна составляет $\pm 0,5$ Да. Сужение этого окна для повышения избирательности фильтра может привести к тому, что целевые пептиды – в силу случайных флуктуаций – не смогут преодолеть фильтрационный барьер, и массив целевых ионов будет недостаточным для получения детектируемого сигнала.

Расширение же окна повышает производительность прибора, однако существенно влияет на достоверность получаемых результатов из-за возникновения интерференции.

Явление интерференции также сужает список пептидов, по которым могут детектированы потенциальные биомаркеры. Так, в работах [Hall и др., 2015; Hüttenhain и др., 2012] список детектированных пептидов сократился более чем в два раза по сравнению с первоначальным.

2.6. *Количественное измерение содержания белков*

Создание масс-спектрометрического метода заключается в синтезе отобранных протеотипических пептидных фрагментов и получении их эталонных масс-спектрометрических измерений. Вместо экспериментального подбора специфических антител селективное распознавание белков производится за счет технических возможностей масс-анализатора. Двойной фильтр пептидных и соответствующих им фрагментных ионов, а также узкие диапазоны регистрируемых при фильтрации соотношений массы к заряду обеспечивают высокую селективность. Поэтому метод мониторинга диссоциативных переходов рассматривается как альтернатива иммуноферментному анализу белков.

Необходимость количественного анализа содержания белка в биологическом образце обусловлена представлениями о том, что белки обладают диагностической значимостью. Связь между содержанием белков и стадией развития установлена для онкологических [Srebrow, Kornblihtt, 2006], сердечно-сосудистых [Garcia-Blanco, Baraniak, Lasda, 2004; Nedelkov, 2008; Sarkozy и др., 2009; Yip и др., 2008], и нейродегенеративных [Cummings, 2005; Garcia-Blanco, Baraniak, Lasda, 2004] заболеваний.

Существует несколько подходов количественного измерения белков в биологическом материале с применением метода мониторинга диссоциативных переходов. Наиболее распространенным является сопоставление интенсивностей пиков изотопно-меченых пептидных стандартов с пиками эндогенных пептидов [Jaquinod и др., 2012]. Выбранный изотопно-меченый пептид (с тяжелыми изотопными метками ^2H , ^{13}C , ^{15}N или ^{18}O) предварительно анализируют в буферных растворах. Основываясь на результатах предварительного анализа или прогноза теоретически наиболее интенсивных переходов, исследователь формирует список отобранных переходов. Далее к трипсинолизированным биологическим образцам добавляют изотопно-меченый пептидный стандарт и проводят масс-спектрометрический анализ образца. По соотношению интенсивностей пиков эндогенного и соответствующего изотопно-меченого пептида, добавленного в триптическую смесь в известном

количестве, исследователь может определить содержание целевого пептида. Еще одной характеристикой, позволяющей идентифицировать пептид, является время выхода его с колонки. Точная масса пептида и время его выхода с хроматографической колонки создают в совокупности массово-временную метку пептида [Николаев и др., 2012]. Аналогичный подход, основанный на массово-временных метках, успешно применялся в масс-спектрометрии негидролизированных белков. Время может определяться экспериментально или с помощью специализированного ПО на основании данных о гидрофобности пептида [Krokhin, Spicer, 2010].

Изучение корреляции между содержанием белка и площадью пика позволило разработать эффективные методы количественной протеомики, не использующие изотопные метки [Neilson и др., 2011]. Одними из наиболее часто применяемых вариантов безметковых методов являются метод определения «площади под кривой» (в англоязычной литературе – area under the curve, AUC) и метод подсчета спектров (в англоязычной литературе – spectral counting).

Метод «площади под кривой» предусматривает измерение ионной интенсивности пептида в определенном временном окне. Правомочность подхода объясняется тем, что площадь под кривой в хромато-масс-спектрометрическом эксперименте прямо пропорциональна ($R^2 = 0,9978$) содержанию анализируемого пептида в концентрационном промежутке 10 фемтомоль – 100 пикомоль [Bondarenko, Chelius, Shaler, 2002; Chelius, Bondarenko, 2002; Liu и др., 2004]. Однако, существует особенность, на которую необходимо обращать внимание при интерпретации результатов эксперимента. В случае коэлюирования пептидов, распределение сигналов пептидов смеси на длительном временном промежутке может существенно снизить разрешающую способность.

Метод подсчета спектров основан на предположении, что фрагментные ионы пептидов в образце белков будут наблюдаться в масс-спектрах пропорционально зависимости от содержания исходного пептида в образце

[Liu, Sadygov, Yates, 2004]. Оценка содержания белка в образце может быть произведена посредством вычисления индекса PAI (protein abundance index), то есть отношения числа зафиксированных в эксперименте пептидов к числу теоретически возможных [Rappsilber и др., 2002]. Безметковый подход требует особого контроля за вариабельностью интенсивности сигнала, а также анализа влияния интерференции коэлюирующих пептидов. Без использования изотопных меток были детектированы 158 потенциальных биомаркеров рака груди [Whiteaker и др., 2007].

В работе [Ludwig и др., 2012] приведен пример количественного анализа клеточных лизатов без использования изотопных меток. Методика анализа основывается на гипотезе существования наиболее «летучего» иона (в англоязычной литературе – “best flyer” hypothesis). Предполагается, что пептид, которому соответствует наиболее интенсивный масс-спектрометрический сигнал в рамках одного эксперимента, будет характеризоваться максимально интенсивными пиками в масштабе всего протеома. Наиболее «летучие» ионы могут быть выбраны для каждого белка по результатам панорамных масс-спектрометрических измерений. Так, в исследовании [Ludwig и др., 2012] для каждого белка были отобраны 3 наиболее «летучих» пептида, на основании анализа которых удалось разработать линейную модель, позволяющую проводить количественную оценку. Модель была оптимизирована для анализа лизата микроба *Leptospira interrogans*. Коэффициент корреляции полученного в эксперименте содержания белков с литературными источниками составил 0,88.

2.7. Стандартизация результатов протеомных исследований

Несмотря на многочисленные литературные источники, указывающие на необходимость стандартизации [Percy и др., 2013a], в практике протеомных исследований отсутствует понятие технологического процесса. По-видимому, это связано с представлениями о протеомике в большей степени как о науке, нежели чем о технологии. При использовании направленного подхода возникает необходимость в создании стандартизированной последовательности операций (протокола), предусматривающей многократную повторяемость.

Протеомное сообщество разрабатывает подходы к стандартизации как экспериментальных протоколов, так и интерпретации полученных результатов [Rai, 2005]. Стандартизация этапов исследовательского процесса позволит снизить общую вариабельность и противоречивость результатов, а также позволит выявить степень влияния каждой операции на конечный результат. Как было показано выше на рисунке 1, направленный масс-спектрометрический эксперимент включает в себя следующие этапы: (1) пробоподготовка; (2) планирование эксперимента; (3) экспериментальная часть; (4) интерпретация полученных результатов.

С точки зрения планирования эксперимента, принципиальное значение имеет количество протеотипических пептидов и их фрагментов, использованных для измерения содержания белка. В работе [Ludwig и др., 2012] рекомендуется отбирать пептиды, которым соответствует наиболее интенсивный сигнал. По результатам этого исследования было получено, что 85% таких пептидов могут быть детектированы с помощью метода МДП. Авторы рекомендуют исследовать от одного до четырех пептидов одного белка, для каждого из которых детектировать от 1 до 6 фрагментов. При этом авторы говорят о высокой корреляции между результатами, полученными по разному количеству пептидов и фрагментов для одного и того же белка. Однако, эти рекомендации, основанные на модели, построенной Ludwig и соавторами [Ludwig и др., 2012], отличаются от рекомендации использовать не менее двух пептидов, изложенной в работе [Carr и др., 2014]. Наилучший результат с наименьшим количеством ошибок, по их мнению, может быть достигнут при использовании двух пептидов, для каждого из которых детектируется две наиболее интенсивные транзиции. В работе [Ludwig и др., 2012] было показано, что использование большего количества диссоциативных переходов увеличивает вероятность возникновения ошибки при детектировании белка. Большее количество входной информации должно было бы приводить к получению более точной информации о белке. Однако, некоторые транзиции,

по-видимому, не информативны, использование транзиций с низкой интенсивностью могут приводить к неточной детекции белков.

Интерпретация результатов измерений, в первую очередь, направлена на оценку воспроизводимости измерений в технических повторах. Исследование, описанное в работе [Addona и др., 2009], заключалось в сравнении результатов анализа 10 протеотипических пептидов семи белков плазмы человека, проведенного методом мониторинга диссоциативных переходов в восьми независимых лабораториях. На основании сравнения результатов был установлен межлабораторный коэффициент вариации, который составил 25%. Данный результат является скорее исключением из правил, поскольку в других работах вариация существенно выше [Hall и др., 2015]. В связи с потребностью получать точные и воспроизводимые количественные данные при верификации и валидации кандидатных биомаркеров требуется создание стандартизированных протоколов получения и обработки результатов измерения содержания белков методом мониторинга диссоциативных переходов [Carr и др., 2014].

В работе [Carr и др., 2014] выдвинуты предпосылки к стандартизации исходя из перспективы применения метода мониторинга диссоциативных переходов в клинической диагностике. Согласно требованиям FDA, диагностический метод не может быть зарегистрирован, если коэффициент вариации измеряемой величины превышает в независимых повторениях 15% [Anderson, 2010]. Для подавляющего большинства протеомных исследований этот порог до сих пор не преодолен: в клинической практике отсутствуют масс-спектрометрические методы измерения диагностических биомаркеров. Можно сделать вывод, что либо измеряется нерелевантная величина, либо величина измеряется с существенной ошибкой в силу человеческого фактора. В первом случае, вместо сигнала, относящегося к определенному, действительно связанному с заболеванием белку измеряется сигнал от другой биомолекулы, возможно даже небелковой природы. Во-втором случае, точность измерительного метода настолько высока, что в результирующий

сигнал человеческий фактор (подготовка пробы лаборантом и/или действия оператора хромато-масс-спектрометрической установки) вносит более существенный вклад, чем события, обусловленные молекулярно-биологическими причинами [Percy и др., 2013a]. Также можно предполагать, что чувствительности аналитических методов недостаточно, чтобы зарегистрировать биомаркерные молекулы в сверхнизких концентрациях [Archakov и др., 2012].

В работе [Carr и др., 2014], резюмирующей опыт в проведении протеомных экспериментов представителей ведущих мировых исследовательских групп, вводится трехступенчатая классификация исследований, проводимых с использованием целенаправленной масс-спектрометрии. Авторы предлагают использовать критерии качества результатов экспериментов в зависимости от целей исследования. На этапе «разведочных исследований», масштаб которых охватывает до 100 аналитов, предлагается допускать низкую точность измерений (коэффициент вариации в технических повторах более 35%) [Berna и др., 2008]

Следующий этап количественного исследования (а также детекция пост-трансляционных модификаций) допускает вариабельность в технических повторах от 35% до 20% [Addona и др., 2009; Addona и др., 2012; Hüttenhain и др., 2012; Keshishian и др., 2009]. И, наконец, клинические исследования предполагают небольшое число аналитов, измеряемых с коэффициентом вариации не более 25% [Picotti и др., 2010; Sherrod и др., 2012; Tang и др., 2011; Zhang и др., 2011].

Изложенный в работе [Carr и др., 2014] подход приемлем для поиска биомаркеров, когда из большого количества потенциально значимых соединений в итоге измеряется несколько. Для хромосомотричного исследования (см. далее) такой подход неприемлем, так как характеризуется низкой производительностью.

Для контроля качества экспериментов, проведенных методом мониторинга диссоциативных переходов, и построения калибровочных кривых

в работе [Percy и др., 2013b] использовали смеси изотопно-меченых пептидных стандартов, масс-спектры которых не интерферируют с масс-спектрами других компонентов биологического образца. Применение стандартов позволило снизить коэффициент вариации в экспериментах по определению содержания 22 белков недеплетированной плазмы человека в шести различных лабораториях на различных приборных платформах до 8% [Percy и др., 2013b].

Очевидным недостатком метода мониторинга диссоциативных переходов является отсутствие статистических моделей оценки достоверности детекции пептидных последовательностей. В панорамных экспериментах такие модели, реализованные в виде программ поиска и сопоставления масс-спектров по базам данных (Mascot, X!Tandem) и оценивающие вероятности идентификации фрагментных ионов, в значительной степени позволяют автоматизировать идентификацию пептидов. Панорамный метод изначально задумывался как высокопроизводительный, поэтому методика потоковой обработки результатов развивалась одновременно с развитием самой технологии. Направленная же протеомика изначально предусматривала ручную обработку результатов, полученных для ограниченного набора белков. К настоящему моменту не существует отработанных подходов для оценки достоверности результатов, полученных направленным анализом. В отсутствие общепризнанных статистических стандартов, адаптированных для направленных экспериментов, решение о достоверности каждого масс-спектра выносится на основании экспертного мнения.

2.8. Программные системы для разработки и анализа результатов масс-спектрометрических экспериментов

С технической точки за последние 10 лет возможности масс-спектрометрических протеомных исследований выросли от идентификации нескольких сотен белков в дрожжах [Godoy de и др., 2008] до практически полного покрытия [Kim и др., 2014; Schaab и др., 2012; Wilhelm и др., 2014b] протеома человека. Увеличение количества идентифицируемых белков потребовало развития биоинформатических методов интерпретации результатов

измерений, в первую очередь, связанных с потоковой обработкой получаемой информации.

Разработаны программные пакеты, позволяющие интегрировать стадии исследовательского процесса от планирования эксперимента до интерпретации результатов. Среди интегрирующих программ можно выделить кросс-платформенные программные пакеты (универсальные), среди которых MRMAid [Cham и др., 2010], MaRiMba [Sherwood и др., 2010], MRMer [Martin и др., 2008], TIQAM [Lange и др., 2008c]. Также существуют платфомерно-зависимое ПО производителей масс-спектрометрического оборудования: MRMPilot (Applied Biosystems), Optimizer (Agilent Technologies) и другие. Последовательно объединяя стадии работы от дизайна эксперимента до статистической обработки результатов, такие программные пакеты позволяют контролировать ход протеомного исследования. Некоторые из них сопровождают весь процесс эксперимента, другие ориентированы прежде всего на дизайн, т.е. создание протокола работы, удобное извлечение полученных данных, поддержку в дальнейшей ручной обработке и публикации результатов. Наиболее широко используемый программный пакет Skyline [MacLean и др., 2010] позволяет интегрировать ранее полученные результаты панорамных измерений, т.е. есть тандемные масс-спектры, в исследовательский процесс направленного протеомного исследования.

Для сопровождения и автоматизации обработки результатов экспериментов, проведенных методом мониторинга диссоциативных переходов, разработан программный пакет mQuest/mProphet, применяющий вероятностные алгоритмы для оценки достоверности идентификации и количественного анализа целевых пептидов. В основе пакета лежит использование заведомо отсутствующих переходов, добавленных в масс-спектрометрический метод в качестве отрицательного контроля [Reiter и др., 2011]: сопоставление сигналов заведомо отсутствующих и целевых переходов позволяют сделать вывод о достоверности получаемых результатов.

Другая интегрирующая платформа – MRMer – разработанная для управления сложными экспериментами, служит для количественного анализа и позволяет сравнивать пары эндогенных и изотопно-меченых пептидов. В программе MRMer используется информация из файлов с отчетами масс-спектрометра, закодированными в не зависящем от платформы mzXML-формате. Анализируя загруженный файл масс-спектра в стандартном формате mzXML, программа автоматически выявляет и группирует пары пептидных и соответствующих им фрагментных ионов, вычисляет интегральные ионные интенсивности и генерирует наглядное представление контроль анализа массива данных, превышающих 1000 переходов.

В литературе описаны масштабные проекты, полностью сопровождаемые одним или несколькими интегрирующими программными пакетами. Например, для разработки требований к воспроизводимости межлабораторных исследований [Abbatiello и др., 2013] был использован пакет Skyline [MacLean и др., 2010]. Skyline обеспечивает теоретический прогноз и отбор наиболее интенсивных переходов из библиотеки спектров, основываясь на созданной по результатам предварительных панорамных экспериментов. Пакет осуществляет удобный перенос выходных данных приборов с тройным квадруполом от различных производителей (Agilent, Applied Biosystems, Thermo Fisher Scientific и Waters). Результаты направленного анализа пакет Skyline позволяет разместить в публичном репозитории Panorama.

2.9. Репозитории результатов масс-спектрометрических экспериментов

Накапливаемый массив масс-спектрометрических данных является ценным источником получения знаний о молекулярных механизмах функционирования живых систем [Wruck, Peuker, Regenbrecht, 2014]. Массив в скрытом виде содержит сведения о свойствах и структуре белковых продуктов, возможных модификациях, а также связи с патологическими процессами.

К такому типу ресурсов относится база данных UniProtKB. Для каждого белка в ней содержатся сведения о наличии опубликованных экспериментальных данных по детекции его протеомными, транскриптомным

или биоинформатическими методами. На основе UniProt в Швейцарском институте биоинформатики (SIB) был создан специальный ресурс neXtProt – онлайн платформа, агрегирующая сведения о человеческих белках [Gaudet и др., 2015]. Источником большей части информации в neXtProt является база данных UniProtKB Swiss-Prot, которая дополняется из других источников, в том числе PROSITE [Sigrist и др., 2009], ENZYME [Bairoch, 2000], STRING [Szklarczyk и др., 2011] Swiss-Model Repository [Kiefer и др., 2009].

В результате выполнения протеомных исследований, использующих масс-спектрометрические подходы, генерируются неоднородные данные о содержании белков в различных типах анализируемых образцов. Массивы данных гетерогенны и, зачастую, противоречивы, что осложняет их интерпретацию. Для выполнения крупномасштабных протеомных проектов требуется сведение воедино накопленных результатов и создание общего принципа стандартизации данных [Orchard, Hermjakob, 2008].

Для хранения и доступа к результатам крупных протеомных проектов создан ресурс MaxQB (<http://www.biochem.mpg.de/maxqb>), позволяющий анализировать и сравнивать полученные высококачественные сведения о содержании белков. Ресурс открывает доступ к результатам количественных протеомных исследований 11 различных клеточных линий, представляющих более половины всех генов человека [Geiger и др., 2012].

База данных Plasma Proteome Database (<http://www.plasmaproteomedatabase.org>), созданная по инициативе международной протеомной организации HUPO, содержит сведения о белках, детектированных в плазме и сыворотке крови здоровых людей с указанием использованного экспериментального метода [Nanjappa и др., 2014] (иммунноферментным анализом, радиоиммунными и масс-спектрометрическими подходами (в том числе методом мониторинга диссоциативных переходов)). Для содержащихся в базе данных значений концентраций белков приводятся ссылки на литературные источники. В мае 2015 года в базе данных содержались данные о 10 546 белках, детектированных в плазме крови здоровых людей

(опубликованы в 209 статьях), из них 279 белков были детектированы методом мониторинга диссоциативных переходов. Сведения о концентрациях приведены для 1278 белков.

Репозитории, предоставляющие открытый доступ к результатам протеомных исследований, стимулируют перекрестный анализ экспериментальных данных, облегчают процесс рецензирования и внедряют в работу исследователей новые программные продукты. Примером такого репозитория является база данных PeptideAtlas, созданная более 10 лет назад [Desiere и др., 2005; Desiere и др., 2006; Farrah и др., 2013] и превратившаяся к настоящему моменту в обширный каталог результатов панорамных масс-спектрометрических экспериментов. Сведения, хранящиеся в PeptideAtlas, представляют принципиальную значимость для создания протоколов направленных масс-спектрометрических экспериментов [Deutsch, 2008]. Сведения о целевых протеотипических пептидах, а также панорамные спектры этих пептидов, доступные из репозитория, используются для планирования направленных протеомных исследований. Для разработки методов мониторинга диссоциативных переходов в рамках базы PeptideAtlas реализован компонент SRMAtlas [Kusebauch и др., 2014]. Посредством третьего компонента, PeptideAtlas SRM Experiment Library (PASSEL), исследователи могут размещать в общем доступе и визуализировать обработанные результаты, полученные методом мониторинга диссоциативных переходов. Размещение экспериментальных данных в PASSEL является одним из обязательных условий при публикации статьи о работах, выполненных в рамках проекта «Протеом человека», в рецензируемых журналах с высоким импакт-фактором. Однако несмотря на это, количество экспериментов, доступных для поиска в PASSEL не так велико.

[Nelson, 2014] [Sharma и др., 2014] результаты эксперимента и шаблон метода, готового для загрузки в масс-спектрометр. Результаты и протоколы экспериментов, хранящиеся в архиве Panorama, представляют собой открытую всего научного сообщества базу знаний направленных протеомных

исследований [Sharma и др., 2014]. На момент февраля 2015 года архив Panorama включал 998 проектов, управляемых 97 лабораториями.

В ресурсе PRIDE³ (Protein Identification Database) [Jones и др., 2006; Jones и др., 2008; Martens и др., 2005] собраны масс-спектрометрические данные, полученные панорамными методами. Ресурс содержит более 13 тысяч протеомных экспериментов, результаты которых могут быть использованы при планировании направленного масс-спектрометрического эксперимента, т.е. для выбора протеотипических пептидов и соответствующих им наиболее интенсивных диссоциативных переходов. Автоматический отбор диссоциативных переходов на основе экспериментов из ресурса PRIDE был реализован в алгоритме программы MRMmaid [Fan и др., 2012]. Математическая модель, использующаяся в этой программе, позволяет производить отбор переходов с наибольшей вероятностью детекции.

Расширенный алгоритм отбора диссоциативных переходов из UniProtKB, NCBI's dbSNP, ExPASy, PeptideAtlas, PRIDE и GPMDB реализован в ПО PeptidePicker [Mohammed и др., 2014].

2.10. *Использование метода мониторинга диссоциативных переходов в экспериментально-клинических исследованиях*

Метод МДП используется для узконаправленных исследований небольших групп белков, объединенных или молекулярно-биологической функцией, или ассоциацией с патологическим процессом [Surinova и др., 2013]. По данным базы PASSEL, в опубликованных исследованиях, проведенных методом мониторинга диссоциативных переходов, было детектировано от 1 до 500 белков. Из 20 опубликованных работ, в которых изучали белки человека, только шесть из них могли считаться высокопроизводительными, так как в них было проанализировано более 50 белков. В среднем различие между количеством белков, отобранных для анализа, и количеством детектированных белков экспоненциально увеличивалось с увеличением группы отобранных для анализа белков. Наблюдалась тенденция сокращения количества

³ URL: <http://www.ebi.ac.uk/pride>

детектированных белков при увеличении числа анализируемых пептидов. Наиболее масштабные работы с использованием метода мониторинга диссоциативных переходов были проведены Huettenhain и соавторами [Hüttenhain и др., 2012]. Проанализировав 1261 белок, отобранный в работе [Anderson и др., 2004] как ассоциированный с раком, Huettenhain и соавторы детектировали 182 белка в деплетированной плазме, динамический диапазон концентраций которых составил 5 порядков, а средняя концентрация - 10 нг/мл. В моче было детектировано 408 белков, ассоциированных с раком. Кроме того, в данной работе было показано, что количественные измерения могут воспроизводиться для 34 белков в 84 образцах плазмы крови. Данная работа является наиболее масштабным опубликованным исследованием с использованием метода мониторинга диссоциативных переходов .

В работе [Mörtstedt и др., 2015] из 246 проанализированных белков в носовой жидкости было детектировано 244, из которых 175 были детектированы с низкой вариабельностью в технических повторях. Из отобранных изначально для анализа 185 белков печени в работе [Chen и др., 2014] были детектированы только 57.

Функциональная возможность метода мониторинга диссоциативных переходов, позволяющего производить точный количественный анализ целевых белков в серии биологических образцов, делает этот подход оптимальным для статистически достоверной верификации биомаркеров. Традиционно для решения этой задачи использовались антительные подходы, однако преимущества масс-спектрометрического метода мониторинга диссоциативных переходов в мультиплексности, селективности и рентабельности, а также высокая корреляция с иммунохимическими методами делают его наиболее предпочтительным для исследования единичных биомаркеров в биологических образцах [Yocum и др., 2008].

Проблема экспериментов, выполненных методом мониторинга диссоциативных переходов по верификации биомаркеров, связана с зачастую недостаточно низкой границей определяемых концентраций целевых молекул в

сложной биологической матрице, в которой содержания белков колеблются в пределах 12 порядков. Дополнительные операции по обогащению пробы целевыми белками могут снизить производительность метода, однако не лишают его однозначных преимуществ в тех случаях, когда требуется исследовать модифицированные белки или, к примеру, подходящие для анализа методом ELISA антитела недоступны или непредставительны. Без дополнительного обогащения нижний предел количественного анализа белков трипсинолизированной плазмы крови составляет примерно 1 мкг/мл с коэффициентом вариации менее 20% [Kuzyk и др., 2009]. К сожалению, содержание важных с клинической точки зрения биомаркеров в плазме крови обычно на три порядка ниже, чем предел обнаружения, которым характеризуются эксперименты, выполненные методом мониторинга диссоциативных переходов в нефракционированной плазме [Anderson, Anderson, 2002]. Приемы предварительного фракционирования и обогащения позволяют проводить представительные измерения низкокопийных биомаркеров. К примеру, технология иммуноаффинной деплеции высококопийных белков плазмы в сочетании с ионообменной хроматографией, использованная в исследованиях [Keshishian и др., 2009] и [Fortin и др., 2009], позволяет понизить предел определяемых концентраций до 1 нг/мл. В этих работах метод мониторинга диссоциативных переходов был применен для определения 9 известных маркеров сердечно-сосудистых заболеваний исследованиях [Keshishian и др., 2009] и специфического антигена простаты [Fortin и др., 2009]. Для достижения такого предела определяемых концентраций в исследовании, направленном на определение потенциального биомаркера некроза сердечной мышцы, перед проведением эксперимента проводили иммунную преципитацию целевого белка [Berna и др., 2007].

Таким образом, практически все исследования, выполненные методом МДП, направлены на поиск биомаркеров патологических процессов.

2.11. Хромосомоцентричный подход

Хромосомоцентричный проект “Протеом человека” объединил усилия научных лабораторий более 20 стран-участниц [Omenn, 2014]. Цель проекта состоит в определении содержания продуктов экспрессии генов в организме человека. Предполагается, что инвентаризация белков всех тканей и жидкостей человека позволит разработать принципиально новые подходы для ранней диагностики и контроля лечения заболеваний.

Трудности реализации проекта прежде всего связаны со сложностью объекта исследования – протеома человека. Биологические жидкости и более 200 типов клеток характеризуются разными наборами белков в разных концентрациях [Kelleher, 2012]. Исследование протеома осложняется вследствие многообразия форм белков (протеоформ), возникающих в результате процессов альтернативного сплайсинга, однонуклеотидного полиморфизма и пост-трансляционных модификаций [Shargunov и др., 2014]. Недостаточная чувствительность традиционных протеомных методов ограничивает количество низкокопийных белков, которые потенциально можно детектировать.

Метод мониторинга диссоциативных переходов отличается высокой чувствительностью, однако обладает ограниченной производительностью, позволяя детектировать не более 100 пептидов за один цикл работы масс-спектрометра. Использование направленного подхода для анализа массива белков, составляющих протеом, предусматривает разделение объемов исследований по хромосомоцентричному принципу. В таком формате результаты анализа белков картируются на соответствующие им белок-кодирующие гены определенной хромосомы. Подход предусматривает разобщенное выполнение типовых операций с последующей интеграцией результатов в едином информационном пространстве [Paik и др., 2012]. Эффективность хромосомоцентричного подхода заключается в том, что отработанные на одной хромосоме технологические стадии можно распространять на другие хромосомы, поскольку все хромосомы примерно

равнозначны по долевному содержанию генов, участвующих в молекулярных процессах и ассоциированных с заболеваниями [Ponomarenko и др., 2012].

В качестве объекта исследования Россией был выбран набор белков, продуцируемых генами хромосомы 18 человека. Российские исследователи детектировали большую часть белков хромосомы 18 человека с помощью метода МДП в трех типах биоматериала: плазме крови, клетках ткани печени, а также в клеточной линии гепатоцеллюлярной карциномы HepG2 [Ponomarenko и др., 2014].

Таким образом, в результате выполнения российской части международного проекта «Протеом человека» был сгенерирован уникальный массив данных, отражающий результаты десятков технологических операций, выполненных цикличным образом на протяжении более трех лет. Операции включали в себя обеспечение условий хранения образцов, подготовку образцов к анализу, расчет предварительных параметров эксперимента, проведение эксперимента, обработку результатов. Накопленная информация может быть использована для проектирования, оптимизации, экспертизы качества исследования протеома.

3. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

3.1. Исходные масс-спектрометрические данные

Исследование результатов масс-спектрометрической детекции белков, кодируемых генами хромосомы 18 человека, произведенной методом мониторинга диссоциативных переходов, осуществляли с использованием экспериментальных данных, полученных в рамках выполнения российской части проекта «Протеом человека» [Zgoda и др., 2013]. Информационный массив представлял собой набор хроматограмм в формате производителя масс-спектрометра (Agilent, США) и включал измерения, проведенные в двух типах биологического материала: в одном образце клеточной культуры HepG2 и одном образце плазме крови здорового человека [Ponomarenko и др., 2014; Zgoda и др., 2013]. Каждому измерению соответствовало от 3 до 5 технических повторов. В соответствии с количеством

первичных хроматограмм анализировали 3,5 тысячи электронных отчетов («CompoundReport»), содержащих результаты обработки масс-спектров. Результаты 568 файлов отражали калибровочные измерения концентрации белков в биологическом материале, остальные файлы содержали результаты количественного анализа белков в биологическом материале. Кроме того, использовали два файла в формате электронных таблиц xlsx, включающих конфигурационные настройки хромато-масс-спектрометрической системы Agilent Triple Quadrupole 6490 (США), в том числе перечень масс-зарядных характеристик фрагментных ионов протеотипических пептидов, использованных для проведения измерений.

Общий объем использованных для выполнения работы исходных данных составил 38 ГБ. Этот объем данных был сгенерирован с использованием хромато-масс-спектрометрической станции под управлением ПО MassHunter Data Acquisition B.04. Первичную обработку данных проводили в ПО MassHunter Data Analysis B.4.0.

Файлы с конфигурационными настройками проверяли на наличие сведений о конфигурации прибора, использовавшейся для мониторинга диссоциативных переходов. К данной группе параметров относились: (1) разрешение на первом и втором квадруполе, которое должно было составлять $\pm 0,7$; (2) допуск на совпадение времени удержания пептидов на колонке должен был составлять ± 1 мин.

3.2. Первичная обработка данных, полученных методом мониторинга диссоциативных переходов

На рис. 2 показаны основные этапы обработки исходных масс-спектрометрических данных. Выбор пиков, соответствующих целевому пептиду, осуществлялся оператором с учетом того, что вершины пиков отдельных диссоциативных переходов должны были совпадать по времени удержания. Также, условием выбора определенных пиков было наличие нескольких (по крайней мере, трех) выраженных, визуально наблюдаемых пиков диссоциативных переходов.

Сглаживание пиков позволяло провести анализ совпадения форм пиков, как описано в работе [Reiter и др., 2011]. Отбирались группы, в которые входили пики диссоциативных переходов, имеющую сходную форму. На практике после визуальной инспекции группы пиков оператор отмечал в ПО MassHunter Data Analysis те диссоциативные переходы, которые относились к одному протеотипическому пептиду (см. рис. 2). Отобранные пики интегрировались, то есть их площади суммировались, чтобы определить интенсивность сигнала родительского пептидного иона. Данная площадь использовалась для оценки отношения сигнала к шуму, при этом интервал для оценки задавался оператором на уровне от трех до пяти значений ширины исследуемого пика. Из этого интервала исключались группы пиков других пептидных ионов, чтобы оценить фоновый шум. Если соотношение превышало 2:1, (как правило, оно превышало 7:1), то пик считался обработанным и экспортировался в электронные отчет.

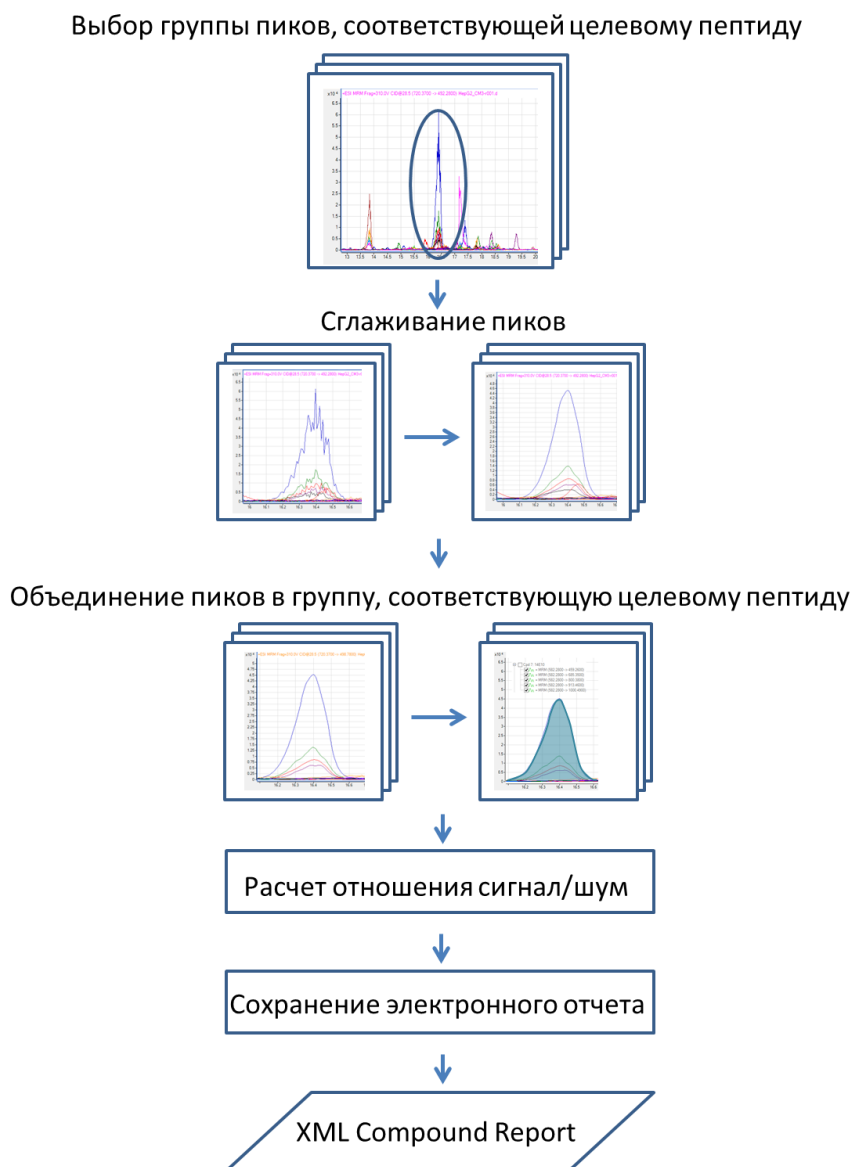


Рисунок 2. Алгоритм обработки исходных масс-спектрометрических данных в ПО MassHunter Data Analysis.

Электронные отчеты, обозначаемые «CompoundReport» в применяемой компанией Agilent номенклатуре файлов данных, анализировали на предмет полноты и непротиворечивости содержащейся информации, как описано ниже (см. раздел 3.3). При сохранении результатов обработки ионных хроматограмм использовали следующие настройки: включение в состав отчета списка пептидных ионов и соответствующих им хроматограмм, а также фрагментных ионов, характеристик и изображений их хроматографических пиков. Электронный отчет

(«CompoundReport») сохраняли в виде папки, содержащей графические файлы с изображениями хроматограмм фрагментных ионов в формате .emf, а также файл с характеристиками аннотированных групп пиков в формате .xml.

3.3. Проверка формата входных данных

Для проверки соответствия исходных данных разработанной структуры была создана вспомогательная программа «CompoundReportConverter.pl». Программу использовали для сопоставления карты фрагментации и конфигурационного метода (.d/192.xml) с электронным отчетом (/Compound Report). При выявлении в составе этих файлов противоречащих сведений или в случае отсутствия необходимых данных файлы считали непригодными для дальнейшего анализа.

Программу «CompoundReportConverter.pl» применяли для проверки наличия информации о результатах этапов обработки файлов (см. предыдущий раздел). Схема электронного отчета представлена на рисунке 3. В файле CompoundReport/Report.xml осуществляли подсчет количества контейнеров <CompoundTable><Compound> и оценивали соответствие их содержимого блоку «ParentIonMass» в карте фрагментации и контейнере <ms1LowMz> в конфигурационном методе. Таким образом, контролировали, что отображенные для исследования пептидные ионы были детектированы на приборе и соответствующие группы пиков были проаннотированы оператором. Проверяли, что в составе группы пиков каждому иону-фрагменту соответствует только один пик, и подсчитывали количество блоков <ChromPeak>, относящихся к каждому пику из контейнера <Chromatogram><ChrompeakList>. Если блоки не соответствовали пикам, то файл считали подлежащим повторной аннотации. Масс-зарядную характеристику каждого фрагмента, указанную в блоке <Chromatogram><ChrompeakList><ChromPeak><BasePeakMz>, сопоставляли со значениями «FragmentIonMass» в составе карты фрагментации. На основании результатов сопоставления выявляли в группе пики фрагментов, не относящихся к выбранному пептидному иону. Для дальнейшей работы проверяли наличие следующих характеристик каждого пика фрагментного иона: абсолютную интенсивность пика, отношение сигнала к шуму пика и время удержания. Наличие

характеристик контролировали по присутствию в блоке <ChromPeak> контейнеров <Height>, <SignalToNoise>, <CenterX>, соответственно. Проведение расчета отношения сигнала к шуму для каждого пика проверяли по наличию в контейнере <ChromPeak> блока <SignalToNoise>.

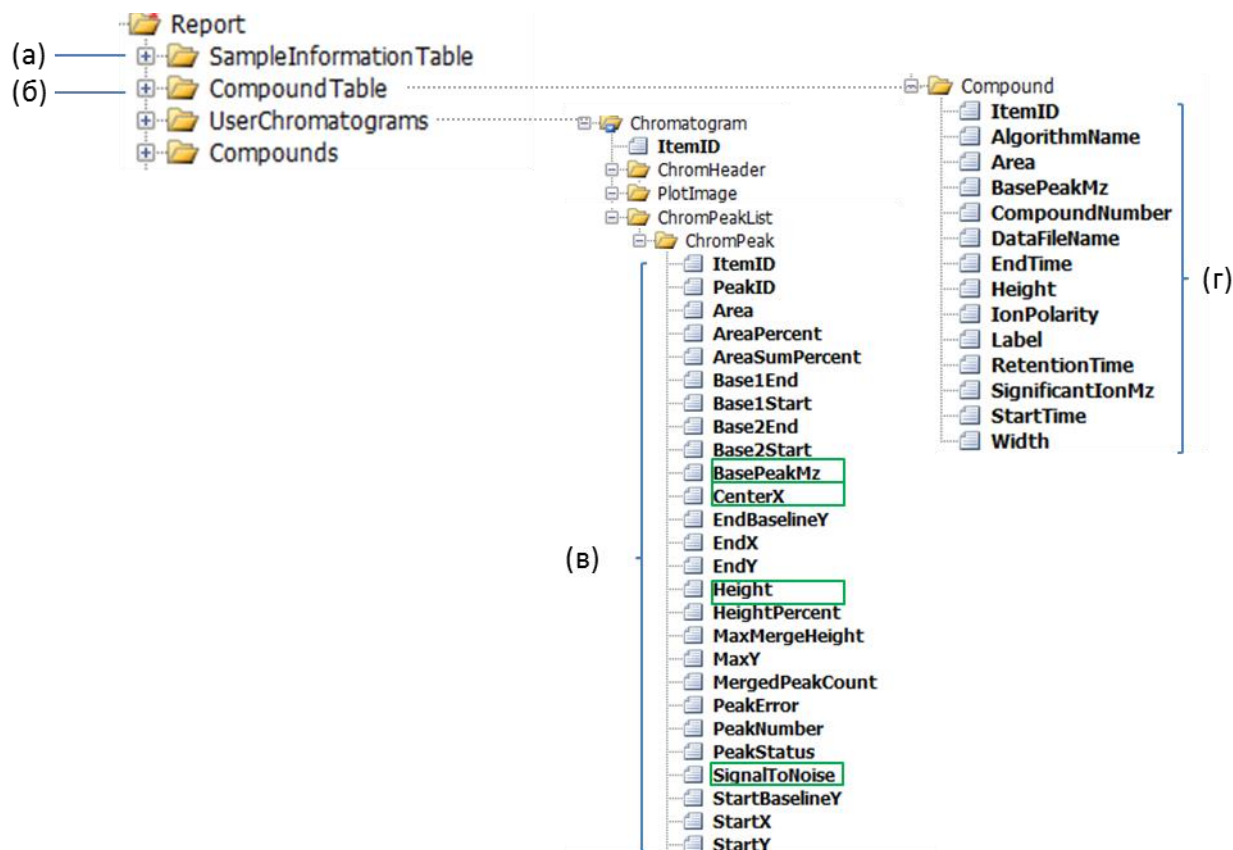


Рисунок 3. Xml-структура электронного отчета о результатах измерения протеотипических пептидов методом мониторинга диссоциативных переходов со следующими выпадающими списками:

(a) *SampleInformationTable* – условия проведения эксперимента;

(б) *CompoundTable* – список пептидных ионов;

(в) *ChromPeak* – характеристики пиков фрагментных ионов: *BasePeakMz* – масс-зарядная характеристика пептидного иона; *CenterX* – время удержания в момент наибольшей интенсивности пика; *Height* – высота пика; *SignalToNoise* – отношение сигнала к шуму;

(г) *Compound* – характеристики пептидного иона.

3.4. Программная реализация базы данных протеотипических пептидов

Программная оболочка для работы с базой данных была реализована компанией ООО «Грамант» (<http://www.gramant.ru>) в виде серверного приложения. Приложение распространяется по лицензии «Creative Commons» в формате веб-

архива (.war), написанного на объектно-ориентированном языке программирования Groovy (релиз 1.6).

Для компиляции приложения использовали набор программных модулей Grails версии 2.0.4, обеспечивающий интеграцию с Java-кодом, и пакетом Java Development Kit в версии 1.6.

В скомпилированном виде программа представляет собой набор упакованных Java-классов. Доступ к базе данных осуществляли посредством Интернет-обозревателя (Internet Explorer 9 и выше, Firefox, Chrome).

Для функционирования программной оболочки на сервере устанавливали следующее программное обеспечение:

- контейнер сервлетов Apache Tomcat версии 7 и выше, реализующий спецификацию сервлетов и спецификацию JavaServer Pages (JSP). Tomcat использовали для запуска веб-приложения в качестве самостоятельного веб-сервера или в качестве сервера контента в сочетании с веб-сервером Apache HTTP Server;
- сервер базы данных (PostgreSQL версии 8 и выше).

Для обеспечения работы программной оболочки доступа к данным использовались следующие дополнительные библиотеки сторонних разработчиков:

- библиотека Apache Shiro для обеспечения аутентификации и авторизации пользователей;
- библиотека Java Mail для отправки электронных оповещений пользователям системы;
- библиотека UniProtJAPI для валидации белков в базе UniProt;
- библиотека FreeHEP VectorGraphics для преобразования формата файлов хроматограмм из emf (Enhanced Metafile) в формат png (Portable Network Graphics);
- библиотека Apache Commons Compress для архивирования-разархивирования файлов экспериментов в формате .d (Agilent);
- библиотека Ajax Uploader в качестве компонента загрузки файлов.

Использование базы данных на персональном компьютере пользователя возможно при наличии веб-обозревателей: Internet Explorer версии 7 и выше,

Firefox версии 3 и выше, Opera версии 9 и выше, Chrome версии 10 и выше, Safari версии 4 и выше.

Веб-совместимую программную оболочку использовали для ввода данных о результатах масс-спектрометрических экспериментов и для подготовки выдач по поисковым запросам и/или по настройкам фильтров. Выдачи обрабатывали с использованием программного комплекса, реализованного на языке программирования Perl. Модули в составе программного комплекса были разработаны в соответствии со структурой данных и обеспечивали автоматизацию выполнения запросов на поиск информации (см. далее раздел 3.6).

3.5. Загрузка результатов масс-спектрометрических экспериментов в базу данных

Результаты масс-спектрометрических экспериментов загружали в созданный информационный ресурс – базу данных протеотипических пептидов, кодируемых генами хромосомы 18. Для создания новой записи об эксперименте использовали раздел **Эксперименты** → **Добавить:** в таблице 1 перечислены поля веб-страницы добавления эксперимента. Страница содержала текстовые поля, в которые вносили название эксперимента, дату его проведения, а также перечень белков и пептидов, подгружаемый из конфигурационного метода настройки масс-спектрометра. При этом формат заполнения поля названия эксперимента был свободным. Для вносимых белков проверялось наличие записей с соответствующим кодом доступа в базе данных UniProt. Все остальные поля заполнялись посредством выбора соответствующего значения из выпадающего списка.

Также вносили информацию об инструменте, на котором проводились измерения, и программном обеспечении, применявшемся для обработки первичных данных. После этого сохраняли внесенные данные в виде черновика. Следующим шагом открывали страницу загрузки первичных масс-спектрометрических измерений, электронных отчетов и конфигурационного метода, использовавшегося для проведения измерений, и вносили их в формате zip.

Таблица 1. Информационные поля для заполнения карточки эксперимента в базе данных протеотипических пептидов с использованием веб-формы.

Наименование поля	Пример значения
Идентификатор эксперимента	156747_Q9P0L0_operational_testing_cycle1
Дата проведения	«7 марта 2013»
Белки и пептиды	Q9P0L0 VAHSDKPGSTSTASFR
Режим снятия спектров	Статический MPM
Аналитическая система	Agilent 6410 QQQ LC-ChipCube
Программное обеспечение	Mass Hunter Qualitative Analysis B 2.0.
Тип пробы	клеточная линия hepg2: 25 million cells
Статус эксперимента	Успешный
Идентификатор PASSEL	https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/PASS_View?identifier=PASS00654

Таблица 1 содержит базовое отображение технологических параметров выполнения операции по получению хромато-масс-спектрометрических данных об одном или нескольких протеотипических пептидах. Форма ввода эксперимента применялась для внесения информации из конфигурационного файла, содержавшего сведения об эксперименте, включая его уникальный идентификатор и дату проведения. Значения этих полей позволяли в дальнейшем отслеживать взаимосвязь между исходными данными, полученными в программном обеспечении Agilent MassHunter DataAnalysis, и записями, внесенными в базу данных протеотипических пептидов, картированных на хромосому 18 человека.

Результаты картирования размещали в поле «Белки и пептиды» (см. табл. 1). При этом, вносили идентификатор белка в номенклатуре UniProt и соответствующие этому белку аминокислотные последовательности протеотипических пептидов. Программная оболочка обеспечивала автоматическую проверку вносимых данных путем сопоставления пептидов с последовательностями белков, входящих в ресурс UniProt.

Поля «Режим снятия ионной хроматограммы», «Название аналитической системы» и «Программное обеспечение» заполняли с целью контроля входных

данных исследовательского процесса. Значения этих полей использовались, чтобы при сопоставлении наборов экспериментов учитывать погрешности, связанные с применением различных модификаций оборудования или программного обеспечения.

Тип пробы, приведенный в таблице 1, вносили в соответствии с контролируемым словарем терминов. Для этого в программной оболочке базы данных был предусмотрен раздел, позволяющий поддерживать терминологическую онтологию. Тип пробы включал как указание источника биоматериала (например, «печень» или «плазма крови»), так и указание способа подготовки пробы к анализу, например, «плазма, после истощения по мажорным белкам». Отдельным типом пробы являлась искусственная проба, полученная в результате смешивания синтезированных пептидов в буферном растворе.

Функция «Добавление пробы» обеспечивало технологическую операцию, когда эксперимент без изменения параметров измерительного оборудования выполнялся в отношении нескольких проб. Такого рода информацию извлекали из базы данных протеотипических пептидов для оценки уровня биологической вариабельности, при этом учитывали даты выполнения эксперимента, чтобы минимизировать аппаратные погрешности.

Статус эксперимента вводился оператором как интегральная субъективная оценка, что результаты эксперимента подтверждают фактическое наличие в пробе целевых протеотипических пептидов. Эксперименты, получившие отрицательную оценку статуса, в данной работе не рассматривали.

3.6. Формирование запросов к базе данных

В работе были использованы формализованные требования для выполнения запросов на формирование выборок из базы данных протеотипических пептидов. Требования обеспечивали извлечение информации в соответствии с взаимоотношениями между сущностями предметной области, установленными структурой данных. Запросы $q(x)$ обеспечивали выдачу следующей информации:

$$q(Exp) = (\{Prot1, Prot2 \dots ProtN\} | \{Pept1, Pept2 \dots PeptN\} | \{HepG2 | Plasma\}) \quad (1)$$

$$q(Prot) = (\{Exp1, Exp2 \dots ExpN\} | \{Pept1, Pept2 \dots PeptN\}) \quad (2)$$

$$q(Pept) = (Prot | (FragMap, \{Run1, Run2 \dots RunN\})), \quad (3)$$

где $Run = \{tr1, tr2 \dots trN\}$; $tr = [ApexRt, Intesity, SNR]$

В формулах (1) – (3) использованы обозначения: Exp – масс-спектрометрический эксперимент; Prot – белок; Pept – пептидная последовательность, входящая в состав аминокислотной последовательности белка; FragMap – карта фрагментации пептида; tr – фрагментный ион; Run – технический повтор; ApexRt – время удержания; Intesity – интенсивность пика фрагмента; SNR – отношение сигнала к шуму; *HepG2*, *Plasma* – биологический материал. Для выполнения одиночных запросов к базе данных использовался функционал программной оболочки; многоуровневые запросы были реализованы в виде набора 0 подпрограмм, написанных на языке программирования Perl.

В работе использовали следующие варианты автоматизированных многоуровневых запросов. Согласно формуле (1), в качестве исходных данных обрабатывали набор идентификаторов экспериментов (или один эксперимент). Результат выполнения запроса включал набор идентификаторов белков (в системе кодификации ресурса UniProt), а также набор протеотипических пептидов Pept1...PeptN. Запрос по эксперименту также включал ключ, позволяющий отбирать данные, полученные на одном или нескольких типах биоматериала (лизат плазмы крови или клеточной линии HepG2).

Второй многоуровневый запрос выполняли в отношении белка или нескольких белков, заданных в виде набора кодификаторов UniProt. Выполнение запроса обеспечивало получение списка экспериментов, в которых производили мониторинг диссоциативных переходов пептидов белка. Также, путем сопоставления аминокислотных последовательностей проверяли уникальность пептида по отношению к определенному белку. Поэтому запрос согласно формуле (2) применяли таким образом, чтобы вместе со списком экспериментов получать

набор протеолитических пептидов, каждый из которых обладал уникальной последовательностью аминокислотных остатков.

Систему запросов использовали для получения информации об отдельном пептиде (см. формулу (3)). В этом случае из результатов выполнения запроса извлекали информацию о белке, к которому относится протеотипический пептид, а также карту фрагментации.

Карту фрагментации подготавливали для каждого пептида с использованием ПО SkyLine. В карту включали диссоциативные переходы, теоретически возникающие в результате коллизионно-индуцированного распада ионов в соударительной ячейке квадрупольного масс-детектора. Для каждого протеотипического пептида для включения в карту фрагментации отбирали 12-15 значений масс-зарядных характеристик дочерних ионов (продуктов диссоциации). Критерии отбора были установлены в соответствии с рекомендациями, приведенными в статье [Martin и др., 2008], включая длину, аминокислотный состав, гидрофобность и другие свойства.

3.7. Оценка результатов масс-спектрометрических измерений

Оценку технической вариабельности проводили для определения погрешностей в режиме работы прибора, в пробоподготовке, а также для планирования эксперимента, в том числе для отбора протеотипических пептидов и их фрагментных ионов. Степень технической вариабельности использовали, чтобы выявить измерения с наибольшей достоверностью и воспроизводимостью, то есть определить, является ли регистрируемый сигнал биологически значимым. Если разница в концентрациях парных пептидов одного белка превышала техническую вариабельность, то анализировали биологически обусловленные причины: альтернативный сплайсинг, одноаминокислотные замены или посттрансляционные модификации.

С использованием функционала веб-интерфейса БД производили экспорт данных в формате электронной таблицы `xlsx`. Экспортированную таблицу конвертировали в формат `csv`, пригодный для автоматической обработки

разработанным пакетом подпрограмм. По результатам обработки рассчитывали воспроизводимость параметров эксперимента.

Воспроизводимость результатов оценивали по коэффициенту вариации следующих параметров: время удержания пептида на колонке (RT), интенсивность фрагментных ионов (Int), отношение уровня сигнала (площади хроматографического пика) к уровню шума по стандартной формуле:

$$CV = \frac{\sigma}{\bar{X}} \times 100\% \quad (4)$$

Профиль фрагментных ионов пептида оценивали, как показано на рисунке 4. Фрагментные ионы сортировали в порядке убывания их интенсивностей, то есть первый ион - самый интенсивный, затем идет ион №2 и ион №3, как показано на схеме.

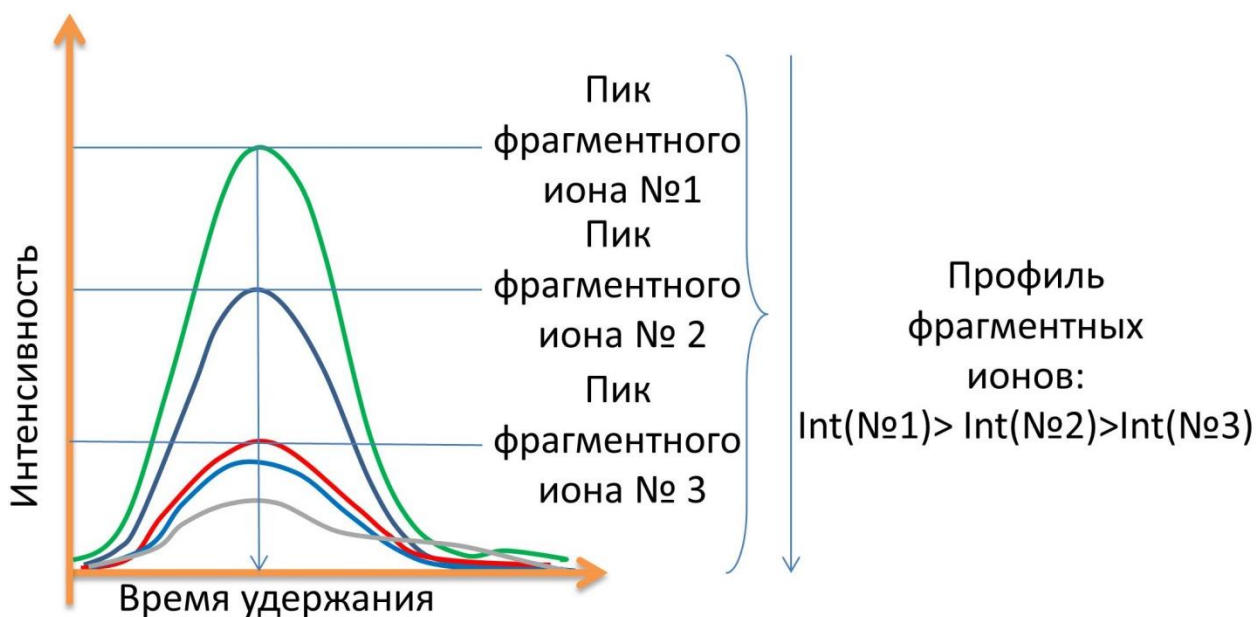


Рисунок 4. Группа хроматографических пиков фрагментных ионов, возникающих в результате индуцированной диссоциации пептидного иона в соударительной ячейке масс-спектрометра. $Int(x)$ – интенсивность масс-спектрометрического сигнала, оцениваемая как высота пика.

В повторном измерении анализировали, сохранился ли порядок следования ионов. Коэффициент вариации рассчитывали как число совпавших по порядку

следования ионов N_1 к их общему числу N с учетом ранее изложенных допущений на приборную погрешность (см. далее раздел « Специализированное программное обеспечение»):

$$CV = \frac{N_1}{N} \times 100\% \quad (5)$$

Оценку воспроизводимости результатов измерений проводили после сегментирования массива экспериментальных данных. Сегментирование заключалось в распределении белков по группам таким образом, чтобы концентрации двух уникальных парных пептидов, относящихся к одному целевому белку, различались не более чем в два раза (группа 1), либо совпадали в пределах одного, двух или более порядков (группы 2-4, соответственно). Содержание целевого белка выражали как отношение количества копий молекулы уникального пептида на клетку линии HepG2. В случае, если измерение концентрации белка проводилось по двум и более пептидам, содержание целевого белка выражали как усредненное значение количества копий пептидов.

Оценку биологической вариабельности проводили путем исследования непротиворечивости данных: непротиворечивостью считали совпадение результатов измерения копийности двух парных пептидов, т.е. пептидов, принадлежащих одному и тому же белку.

С использованием базы данных протеотипических пептидов находили парные пептиды, т.е. пептиды, аминокислотные последовательности которых соответствуют одному белку. Для парных пептидов проводили контроль, что методические условия выполнения измерительных экспериментов, выполненных в отношении каждого пептида из пары, являются эквивалентными. Выделяли массив непротиворечивых данных, т.е. данных, в которых значения концентрации в одном биоматериале для пептидов одного белка различались менее чем на один порядок. Массив данных, в отношении которых наблюдались противоречия, включал белки, для которых различия в результатах измерения между парными пептидами превышали один порядок. Результаты измерений выражались в количестве копий пептида в одной клетке клеточной линии HepG2 или в 1 мл плазмы крови человека.

3.8. Используемые информационные ресурсы

Для получения данных о продуктах генов хромосомы 18 на транскриптомном и протеомном уровнях, их локализации в клеточных компонентах, а также вовлеченности в молекулярно-биологические процессы и механизмы развития заболеваний, использовали БД UniProt⁴.

Для каждого гена хромосомы 18 из БД UniProt загружали сведения о детекции его белкового продукта экспериментальными протеомными методами или данные об экспрессии белок-кодирующего гена. По данным UniProt на апрель 2015 года, экспрессия 201 гена хромосомы 18 была подтверждена на протеомном уровне, 53 генов – на транскриптомном, три гена были предсказаны по гомологии, а остальные имели неопределенный статус.

Аминокислотные последовательности вариантов всех аннотированных в базе данных белков, образованных в результате альтернативного сплайсинга, также загружали из UniProt⁵ и подвергали виртуальному трипсинолизу. Уникальность последовательности каждого детектированного протеотипического пептида белка хромосомы 18 проверяли по отношению к каноническим формам и сплайс-опосредованным вариантам белков в масштабе всего протеома.

Результаты масс-спектрометрических измерений белков, кодируемых генами хромосомы 18, сравнивали с данными экспериментов, выполненных другими исследовательскими группами. Базу данных Plasma Proteome Database (PPD)⁶ [Nanjarra и др., 2013] использовали для получения сведений о детекции и количественном измерении белков масс-спектрометрическими методами (включая метод мониторинга диссоциативных переходов и панорамный подход) в плазме крови здоровых людей. Из базы данных PPD загружали список пептидов, по которым были детектированы масс-спектрометрическими методами белки в плазме крови человека⁷, список белков, детектированных методом мониторинга

⁴URL: <http://www.uniprot.org/uniprot/>

⁵URL: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot_varsplic.fasta.gz

⁶URL: <http://www.plasmaproteomedatabase.org/>

⁷URL: http://www.plasmaproteomedatabase.org/PPD_download/peptide_for_experimental_evidence_020215.xls

диссоциативных переходов⁸, а также доступные из литературных источников сведения об измеренных концентрациях белков в плазме здоровых людей⁹.

Поиск сведений производили по идентификаторам генов хромосомы 18 (UniProt AC) и последовательностям протеотипических пептидов. Из записей о 10546 белках, содержащихся в базе PPD и обнаруженных в плазме крови здоровых людей различными экспериментальными методами (опубликованы в 209 статьях), были отобраны данные о 153 белках хромосомы 18, детектированных методом мониторинга диссоциативных переходов. Сведения о концентрациях в PPD были приведены для 1278 белков, в том числе для 22 белков, кодируемых генами хромосомы 18, из них для трех - установлены масс-спектрометрическими методами.

Информацию о содержании белков в клеточной линии HepG2 получали из базы данных MaxQB, содержащей результаты панорамных протеомных исследований 11 клеточных линий [Schaab и др., 2012]. Всего из ресурса MaxQB загрузили данные о количестве копий белка в одной клетке для 80 белков, кодируемых генами хромосомы 18 человека в клеточной линии HepG2 [Geiger и др., 2012].

Перечень белков хромосомы 18, идентифицированных в панорамных масс-спектрометрических экспериментах в других типах биоматериала, получали из репозитория PRIDE с помощью его RESTful-сервиса (<http://www.ebi.ac.uk/pride/ws/archive/>). В качестве поискового запроса использовали идентификаторы Uniprot AC. Всего загрузили сведения о 266 белках. Количество идентификаций в масс-спектрометрических экспериментах составляло от 1 до 595. Из них в плазме крови было идентифицировано 136 белков хромосомы 18, которые были идентифицированы максимум в 166 экспериментах.

⁸ URL:http://www.plasmaproteomedatabase.org/PPD_download/Multiple_reaction_monitoring.xls

⁹ URL:http://www.plasmaproteomedatabase.org/PPD_download/concentration_020215.xls

3.9. Специализированное программное обеспечение

Для первичной обработки экспериментальных данных использовали программный пакет MassHunter Qualitative Analysis, поставляемый компанией Agilent (США) вместе с хромато-масс-спектрометрической системой Agilent Triple Quadrupole 6490.

В данной работе для обработки масс-спектров, сглаживания пиков, отбора группы пиков, относящихся к одному пептидному иону, и генерации отчетного файла («Compound Report») использовали версии B.02, B.04, B.05 программного пакета MassHunter Qualitative Analysis (см. рисунок 2).

Для поиска обогащенных терминов генной онтологии в списке целевых генов в сравнении с фоновой выборкой генов в масштабе всего генома человека использовали ПО GOrilla. Поиск проводили, загружая в программу два списка белков – «фоновый» перечень белков хромосомы 18 и «целевой» список из 23 белков полученной выборки. Программой производилось сравнение частот встречаемости терминов онтологии генов в двух списках. Результат сравнения был представлен в виде графа с терминами онтологии генов на вершинах. Связи между вершинами устанавливались в соответствии с иерархией онтологии генов.

Контроль соответствия детектированных пептидных последовательностей белкам хромосомы 18, а также соответствия масс-зарядных характеристик пептидных ионов и ионов-фрагментов, использованных для настройки прибора, осуществляли с помощью ПО Skyline 2.6 lkz 64-битной системы (<https://skyline.gs.washington.edu/labkey/project/home/software/Skyline/begin.view>) [MacLean и др., 2010].

Расчет теоретического времени удержания пептида проводили с использованием ПО Theoretical Chromatograph [Gorshkov и др., 2006] для оценки различий времени удержания результирующей выборки протеотипических пептидов белков, кодируемых генами хромосомы 18. Оценку проводили на основании физико-химических свойств пептидов и характеристик хроматографической колонки. В таблице 2 приведены значения введенных параметров, использованных для расчета ожидаемого времени от момента ввода

пробы вещества в хроматограф до момента регистрации максимальной интенсивности хроматографического пика пептида.

Таблица 2. Настройки ПО Theoretical Chromatograph для оценки времени удерживания на основе физико-химических свойств пептидов и характеристик хроматографической колонки

Параметр	Значение
Длина хроматографической колонки, мм	50
Внутренний диаметр хроматографической колонки, мм	2,1
Размер пор сорбента, Å	1,8
Начальная концентрация фазы В, %	0
Конечная концентрация фазы В, %	50
Длительность градиента, мин	45 мин
Время задержки, мин	0,6
Скорость потока, мкл/мин	100
Концентрация ацетонитрила в фазе А, %	2
Концентрация ацетонитрила в фазе В, %	80
Соотношение неподвижной и подвижной фаз:	RP/ CAN+FA

Данные об интерференции пептидных ионов получали с помощью ПО SRM Collider, версия 1.4 [Röst, Malmström, Aebersold, 2012]. В программу загружали перечень протеотипических пептидных последовательностей белков хромосомы 18. При расчете учитывали только фрагментные ионы b- и y-типов. Ширину окна первого квадруполья, фильтрующего пептидные ионы, устанавливали равной 0,7, тогда как ширину окна второго квадруполья, осуществлявшего отбор фрагментных ионов, устанавливали равной 1,0. Минимальное значение масс-зарядных характеристик фрагментов составляло 300, а максимальное 1500, что соответствовало селективности использованного масс-спектрометра [Shi и др., 2012]. Поиск интерферирующих соединений осуществляли по базе данных триптических пептидов, сгенерированной на основании транслированных из генома человека белковых последовательностей. Для каждого пептида все возможные интерферирующие ионы визуализировали в виде графика,

расположенного в координатах m/z пептида – по оси абсцисс/расчетное время удержания – по оси ординат.

Автоматизацию запросов, а также фильтрацию, сегментацию и проверку достоверности данных осуществляли посредством трех комплектов скриптов, написанных на интерпретируемом языке программирования Perl. Скрипты были использованы для исследования параметров, измеряемых в ходе хромато-масс-спектрометрического эксперимента.

Для сопоставления технических повторений по параметрам высоты пиков и соотношения сигнала к шуму на экстрагированных хроматограммах разработали подпрограмму, осуществляющую синтаксический разбор результатов экспорта данных из созданной базы данных протеотипических пептидов <http://www.pikb18.ru/>. Для расчета коэффициента вариации результатов, полученных в нескольких технических повторениях, использовали стандартные функции библиотеки `Statistics::Descriptive ()` для языка программирования Perl.

Наиболее сложной с методической точки зрения задачей являлась оценка воспроизводимости профиля фрагментных ионов. Для иллюстрации этой задачи приводим типовой пример, предположив, что в эксперименте «А» диссоциация пептидного иона привела к возникновению трех фрагментов. Обозначим их номерами по мере убывания интенсивности, т.е. фрагмент №1 имеет наибольшую интенсивность пика, фрагмент №2 – следующую по величине, а фрагмент №3 – наименьшую. В следующем эксперименте «Б» те же фрагменты, отсортированные по убыванию интенсивности пиков, выстроились в другом порядке: №1, №3 и №2. В случае, если различие между интенсивностями ионов №2 и №3 было в пределах приборной погрешности (т.е. <15%), то сочетание диссоциативных переходов №1 → №2 → №3 считали эквивалентным сочетанию №1 → №3 → №2. Если же отличие превышало 15%, то профили фрагментных ионов пептида, полученные в экспериментах «А» и «Б», считали различными. При этом, в качестве экспериментов «А» и «Б» подразумевались как технические повторения в рамках одного эксперимента, так и измерения, выполненные в разных типах биоматериала.

4. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

4.1. Структура данных для описания результатов направленного масс-спектрометрического эксперимента

При разработке структуры данных основывались на том, что в существующих репозиториях для депонирования протеомных данных, как правило, не включено понятие эксперимента как совокупности технических повторений. Эксперимент описывается как сочетание условий проведения и полученных результатов, тогда как эксперимент как серия технических повторений измерений в одинаковых условиях не поддерживается.

Разработанная структура данных включает информационные объекты, позволяющие описать процесс и результаты детекции белков методом мониторинга диссоциативных переходов. Структура создана с учетом того, что высокопроизводительная масс-спектрометрия не позволяет исследовать белки целиком, поэтому в ходе пробоподготовки проводят гидролиз белка на пептиды. Полагают, что результат количественного измерения содержания пептидов в биологическом материале идентичен количественному содержанию белка. Направленная масс-спектрометрия основана на детекции пептидного иона, а также его фрагментов, возникающих в результате электроспрейной диссоциации масс-спектрометра. Сочетание пептидного иона и его фрагментов позволяет точно детектировать пептид в случае отсутствия в анализе посторонних ионов, интерферирующих с сигналом пептида. В разработанной структуре данных каждый фрагмент пептидного иона характеризуется временем удержания на хроматографической колонке, интенсивностью хроматографического пика и отношением сигнала к шуму.

Надлежащей практикой является проведение нескольких технических повторов в рамках одного эксперимента. Интерпретация результатов с целью нивелирования приборных ошибок осуществляется оператором. Таким образом, при разработке структуры данных допускали, что действия оператора при обработке хроматограмм не приводят к искажению информации о детекции пептидов.

На рисунке 5 представлена разработанная структура данных, включающая следующие объекты масс-спектрометрического эксперимента: «белок», «пептид», «фрагмент», «эксперимент», «технический повтор». Источником неточности измерений, проводимых методом мониторинга диссоциативных переходов являются допущения, применяемые при переходе от результатов технических повторов к результатам эксперимента и далее - от результатов измерений фрагментов пептидов к результатам измерений белков. Разработанная структура данных позволяет оперировать информационными объектами для оценки влияния использованных допущений на итоги протеомного исследования.

Объекты структуры данных связаны друг с другом стандартными типами связей: один ко многим и многие ко многим. Пептид как объект созданной информационной структуры данных обладает свойством протеотипичности. Понятие протеотипичности пептида включает свойство «уникальность» – аминокислотная последовательность пептида соответствует канонической (наиболее часто встречающейся, по данным ресурса UniProt) последовательности одного белка, и не встречается в других кодирующих участках генома человека. Если пептид является уникальным и соответствует одному белку, то этот белок является целевым по отношению к пептиду. Если два уникальных пептида относятся к одному белку, то они обозначаются как парные. Кроме того, протеотипический пептид возможно достоверно детектировать с использованием масс-спектрометрического анализа.

В большинстве опубликованных в области направленной протеомики работ используется один целевой пептид [Picotti и др., 2010; Sherman и др., 2009]. Исключение составляют работы группы [Ludwig и др., 2012], в которых при количественных измерениях содержания белка использовали два и более целевых пептида. Количественные оценки содержания белка, полученные по нескольким пептидам, могут существенно отличаться за счет различий в их ионизационных свойствах, которые выражаются в разном уровне масс-спектрометрического сигнала.

В рассматриваемой структуре данных изложенная особенность масс-спектрометрических измерений отражена с использованием особого объекта, «калибровочное уравнение» (см. рис. 5). Он включает в себя несколько экспериментов, заключающихся в измерении интенсивности масс-спектрометрического сигнала в нескольких концентрационных разведениях пептида. На основе серии разведений рассчитываются параметры калибровки – смещение и угол наклона стандартной кривой [Addona и др., 2009].

Один эксперимент может быть связан как с одним, так и со многими белками/пептидами. Пептиды одного белка связаны с разными экспериментами/техническими повторами. Технические повторы одного эксперимента связаны как с разными, так и с одинаковыми фрагментами/пептидами/белками. Если связь белков/пептидов с экспериментами/техническими повторами является результатом планирования эксперимента, то различия между техническими повторами одного эксперимента отражают воспроизводимость полученных результатов. Это ключевое свойство структуры данных позволяет посредством системы запросов отобрать белки, измерения которых не подвержены искажениям в силу технических особенностей масс-спектрометрического метода мониторинга диссоциативных переходов. В то же время, различия между результатами разных экспериментов, в которых проводились измерения парных пептидов одного белка, позволяют посредством мета-анализа строить гипотезы о механизмах функционирования биологических систем.

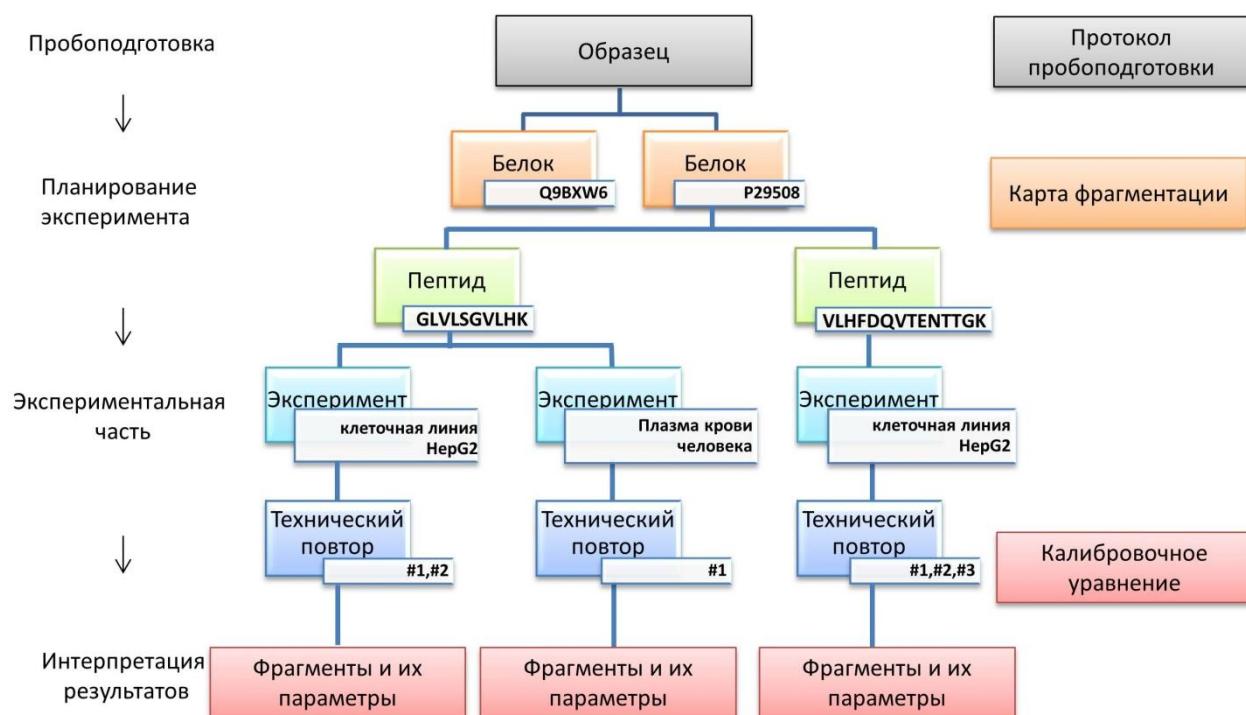


Рисунок 5. Структура данных протеомных исследований, основанных на высокопроизводительном методе мониторинга диссоциативных переходов.

На рисунке 5 видно, что разработанная структура данных описывает этапы исследования протеотипических пептидов масс-спектрометрическим методом МДП. Структура учитывает специфику измерительного инструмента (хроматографической установки с масс-спектрометрическим детектором) и включает блок, описывающий планирование эксперимента и выполнение регламентированных технологических операций с биообразцами.

Связь пептидного иона и иона – фрагмента называют диссоциативным переходом (транзицией). Метод мониторинга диссоциативных переходов основан на детекции пар пептидного иона и иона – фрагмента. Это отражено в разработанной структуре данных. Каждый фрагмент в структуре данных обладает как качественной характеристикой (детектирован/не детектирован), так и количественными: интенсивность хроматографического пика, отношение сигнал\шум, время удержания.

Ограничения существующих структур данных, потребовавшие решения в рамках отдельной задачи данной работы, возможно, связаны с применением лабораторных информационных систем. Разработка таких систем возможна при

высокой степени стандартизации, что проблематично в условиях научного поиска; чаще всего информационные системы оправдывают себя в практике рутинных измерений. В направленной протеомике для случая рутинных измерений применяется программное обеспечение SpectraDive компании Biognosys (Швейцария), и программный пакет SkyLine/Panorama (см. Обзор литературы). Хромосомотцентричное исследование, целью которого является инвентаризация набора белков, не подчиняется логике указанных программных решений именно с точки зрения структуры данных. При исследовании хромосомы речь не идет о многократно повторяющихся измерениях небольшого количества белков в определенном биоматериале. Наоборот, требуется постоянное расширение количества аналитов, вплоть до покрытия всех генов хромосомы.

В данной работе хромосомотцентричный подход был использован как сопрягающее звено между требованиями к стандартизации данных и постоянным изменением условий эксперимента. Такие изменения неизбежно происходят в процессе исследования различных белков. Хромосомотцентричный подход позволяет реализовать гибкую структуру данных, и за счет этого: во-первых, обеспечить контроль над процессом получения новых знаний, а во-вторых, сохранить достаточное количество степеней свободы, необходимых при крупномасштабном научном поиске.

В таблице 3 приводится сравнение разработанной структуры данных с аналогами. Типовая структура, к которой относится большинство поставляемых с масс-спектрометрическим оборудованием программ для обработки хроматограмм, рассмотрена на примере ПО Mass Hunter, поддерживает только одну из четырех взаимосвязей, необходимых для выполнения протеомного эксперимента: связь между соединением и условиями выполнения эксперимента (картой фрагментации).

Таблица 3. Сравнение взаимосвязей между объектами разработанной структуры данных с аналогами, реализованными в пакетах программ общего назначения (на примере MassHunter Data Analysis), и специализированных пакетов для протеомных исследований (SkyLine, PASSEL, SpectraDive).

Взаимосвязь	MassHunter Data Analysis	SkyLine	PASSEL	SpectraDive	Разработанная структура данных
белок-эксперимент	Нет	Да	Да	Да	Да
белок-пептид	Нет	1 ко многим	1 ко многим	1 к 1	1 ко многим
пептид-карта фрагментации	Да	Да	Нет	Нет	Да
эксперимент-технические повторы	Нет	Нет	Нет	Усреднение по техническим повторам	Да
пептид-группа пиков	Нет	Да	Да	Да	Да

В отличие от типового масс-спектрометрического ПО, программы Skyline, Passel, SpectraDive обеспечивают взаимосвязь между белком и его гидролитическими пептидами. В случае программных систем PASSEL [Farrah и др., 2012] и SpectraDive представляет сложность автоматическая обработка информации об условиях эксперимента. Для системы PASSEL сложность связана с чрезмерной гибкостью универсального формата TraML [Deutsch и др., 2012], практически не применимого в условиях конкретного эксперимента. Разработанная в данной работе структура данных является частным случаем подготовки данных о направленном масс-спектрометрическом эксперименте в формате TraML, соотнесенным с хромосомотцентричным дизайном протеомного исследования.

При работе с ПО SpectraDive становится очевидной еще одна проблема, отмеченная в таблице как отсутствие связи пептида с картой его фрагментации. В силу этой особенности применение программы ограничено белками и пептидами, предлагаемыми разработчиками программного обеспечения. Это не позволяет использовать информационную структуру ПО SpectraDive в поисковых задачах

хромосомоцентричного исследования. Фиксированные взаимосвязи между пептидом и его фрагментами несовместимы с поисковым хромосомоцентричным форматом исследования, целью которого является нахождение новых протеотипических пептидов. Хромосомоцентричный формат предусматривает возможность поиска оптимального объекта исследования на этапе планирования эксперимента; результатом такого эксперимента становятся как сведения об оптимальном пептиде и его фрагментах, так и перечень пептидов и фрагментов, непригодных для исследования методом МДП. Связь пептида с картой фрагментации позволяет сохранять сведения о таких пептидах, тем самым предотвращая их повторное использование в новых экспериментах. При этом на сегодняшний день в существующих репозиториях возможность публикации результатов поиска оптимальных пептидов/фрагментов не существует в явном виде.

Из таблицы 3 (колонка SpectraDive) видно, что следствием ограничения на создание взаимосвязей «пептид-карта фрагментации» является то, что взаимосвязь между белком и пептидом устанавливается в соотношении «один-к-одному». Не вызывает сомнений, что в условиях неспецифичного по отношению к последовательности пептида гидролиза логически обоснованной является связь «один белок к нескольким пептидам», а не «один белок – один пептид» (см. рис. 5). Именно сопоставление пары уникальных для белка пептидов является основополагающим для разделения (сегментирования) массива масс-спектрометрических данных на группы (см. далее раздел 4.3).

Выводом из сопоставления разработанной структуры данных с аналогами (Табл. 3) является отсутствие в альтернативных программных продуктах взаимосвязей эксперимента с техническими повторениями измерений. В условиях небольших объемов экспериментальных данных это не является критичным, так как оператор имеет возможность вручную обрабатывать все файлы и интерпретировать полученные результаты. В случае же хромосомоцентричных исследований для 2500 экспериментов с 3–4 техническими повторами количество файлов с результатами уже составляет тысячи. Ручная интерпретация этих данных

не представляется возможной. В ПО SpectraDive, которое является наиболее близким аналогом разработанной структуры, информация по техническим повторениям выдается в усредненном виде, причем только в отношении одного параметра – содержание пептида в пробе биологического образца. Усредненных данных недостаточно для полноценного исследования воспроизводимости между техническими повторами, поскольку отсутствует первичная информация о характеристиках пиков на ионной хроматограмме.

Исследование протеома масс-спектрометрическими методами имеет ряд особенностей, определяемых структурно-функциональными и физико-химическими свойствами белков, а также соответствующих им пептидных последовательностей. В силу этого направленный масс-спектрометрический эксперимент в области протеомики требует разработки формальной структуры данных, основывающейся на объединении ключевых понятий протеомики и масс-спектрометрии.

Существующие структуры данных, реализованные в стандартных программных продуктах и базах данных, не подходят для описания взаимосвязей между объектами исследования. Например, в программном пакете MassHunter, предлагаемом для обработки ионных хроматограмм, центральным объектом структуры данных является соединение (см. рисунок 3). Отсутствует взаимосвязь между пептидом и его родительским объектом – белком, в результате гидролиза которого и образуются пептиды. Ограничение становится существенным при работе в хромосомотричном формате, когда масштабы измерений охватывают сотни и даже тысячи экспериментов.

Существенным отличием от существующих баз данных является то, что разработанная структура данных позволяет хранить сведения о негативных результатах детекции белков в биологическом материале. Публикация этих результатов важна для выявления потенциальных биомаркеров, а так же имеет принципиальное значение при планировании дальнейших экспериментов. Сведения об отрицательных результатах детекции имеют принципиальную

значимость для планирования последующих экспериментов, а также при изучении тканеспецифичности биологических маркеров [Naegle, Gough, Yaffe, 2015].

В созданной структуре данных (см. рис. 5) понятие технических повторов эксперимента является ключевым. В каждом повторении сохраняется полный объем данных, полученных в результате хроматографического разделения пептидной смеси с масс-спектрометрической детектированием состава разделяемых фракций. Для каждого повторения содержатся характеристики хроматографических пиков, включая время выхода с колонки, высоту, площадь, отношение сигнала к шуму, симметрия пика, положение вершины и другие параметры (см. рис. 3). Автоматизированная обработка информации в технических повторениях позволила сформировать результирующую выборку пептидов, которые обладали наибольшей правдоподобностью детектирования и количественного измерения. Далее в работе будет показано, что анализ результирующей выборки, возможный при применении структуры данных и надлежащих интерактивных средств автоматизации к созданной структуре данных, позволяет распределить протеотипические пептиды по нескольким группам.

4.2. Система запросов к информации о протеотипических пептидах

Структура данных, предложенная в настоящей работе, потребовала разработки системы информационных запросов. Запросы были реализованы в соответствии с формулами (1)–(3), приведенными в методическом разделе. Исполнение запросов осуществляли в рамках информационного ресурса с автоматизированным функционалом.

Система запросов была выполнена в виде двух программных пакетов, первый из которых представлял собой объектно-ориентированную интерактивную веб-систему. Функционал этой системы в сжатой форме изложен далее в настоящей главе. Веб-система служила для загрузки и нормализованного представления исходных данных, для исполнения запросов на отбор информации и для экспорта отобранных данных в универсальном формате электронной таблицы CSV (значения, разделенные запятой). Веб-ориентированная система позволяла отбирать эксперименты по условиям их проведения, кодам доступа

детектированных белков в базе данных UniProt, по пептидам и по биологическому материалу (см. формулы (1)-(2)). Экспортированные файлы обрабатывались подпрограммами, запускаемыми из командной строки.

Второй пакет, реализованный с использованием средств скриптового кодирования Perl, в общей сложности включал 3 подпрограммы. Приведенные в таблице 4 сведения отражают последовательные этапы обработки базы данных протеотипических пептидов, предназначенных для количественного измерения содержания белков хромосомы 18.

Таблица 4. Подпрограммы, исполняемые в среде интерпретатора команд Perl для последовательной обработки базы данных протеотипических пептидов.

Наименование подпрограммы	Предназначение	Входные параметры	Выходные данные
TechnicalRepeats.pl	Оценка коэффициента вариации масс-спектрометрических параметров в технических повторях	Значения отношения сигнала к шуму, интенсивности, площади, времени удержания каждого фрагмента пептида в технических повторях эксперимента	Коэффициента вариации для значения отношения сигнала к шуму, интенсивности, площади, времени удержания каждого фрагмента, профиля переходов пептида в технических повторях эксперимента
Standart.pl	Сравнение профиля переходов пептида, зарегистрированного в различных типах биологического образца и/или синтетическом растворе	Профиль переходов пептида в различных биообразцах и/или в синтетическом растворе	Коэффициент вариации профиля переходов
Consistency.pl	Сравнение результатов количественных измерений парных пептидов, соответствующих одному белку	Сведения о результатах количественных измерений пептидов	Соотношение измеренных концентраций парных пептидов

Входными данными для всех подпрограмм служили выборки экспериментов, загруженные из базы данных или полученные после использования одной из подпрограмм. Для каждой подпрограммы при запуске устанавливалось пороговое значение фильтра. Выходные данные каждой из подпрограмм выдавались в формате, аналогичному формату входных данных. Таким образом, обеспечивали возможность последовательного применения подпрограмм.

В соответствии с разработанной структура данных запросы были реализованы для трех объектов предметной области: эксперимент, белок и пептид. Для эксперимента в результате запроса получали информацию о наборе белков, измерения в отношении которых проводились в одинаковых условиях. Поскольку каждому белку соответствовали протеотипические пептиды, то второй вариант исполнения запросов предусматривал получение набора пептидных последовательностей, которые анализировались в одном или нескольких экспериментах. Соответственно, если в качестве запроса использовался идентификатор белка, то результат представлял собой либо перечень экспериментов, либо перечень пептидов.

В работе часто использовали запрос, где в качестве ключа указывался пептид в виде аминокислотной последовательности (см. формулу (3)). Каждый пептид уникальным и однозначным образом должен был соответствовать одному белку, поэтому в качестве выполнения запроса возвращался идентификатор белка. Развернутый запрос по пептиду позволял получить перечень теоретически рассчитанных и экспериментально детектированных фрагментных ионов, а также множество технических повторений, в которых осуществлялись измерения методом мониторинга диссоциативных переходов. Результат запроса раскрывался в форме (см. рис. 7), где каждое техническое повторение представляло совокупность детектированных переходов. Переход, в свою очередь, характеризовался тремя параметрами: время выхода с хроматографической колонки, интенсивность хроматографического пика, соотношение сигнала к шуму.

Для работы с системой запросов были реализованы несколько логических сценариев. Согласно первому сценарию, формировалась выборка экспериментов

или технических повторов, выполненных на одном типе биоматериала. Во втором сценарии создавалась выборка, соответствующая заданному набору целевых белков и соответствующих данным белкам пептидов. Эти два сценария реализовывались посредством встроенного функционала базе данных. Третий сценарий предусматривал создание для запрашиваемого пептида выборки фрагментов, детектированных методом мониторинга диссоциативных переходов. Последний сценарий позволял получить выборку, отвечающую заданному пользователем фильтру на характеристики фрагментов. Изложенная система запросов применялась в дальнейшей работе, чтобы реализовать каскад фильтров и провести сегментирование массива экспериментов.

Разработанная система запросов позволила охарактеризовать результаты загрузки экспериментов в базу данных протеотипических пептидов. Следует учитывать, что объектами загрузки являлись эксперименты, несмотря на то, что ключевым объектом базы данных являются протеотипические пептиды. После загрузки в базу данных содержались сведения о 1689 экспериментах, выполненных масс-спектрометрическим методом мониторинга диссоциативных переходов. В состав экспериментов вошли данные о 275 белках, закодированных на хромосоме 18 человека, в том числе были данные о результатах измерений в гидролизатах клеточной линии и плазме крови, соответственно. В совокупности для всех экспериментов была загружена информация о 2247 протеотипических пептидах. Более чем для половины пептидов было детектировано по 5 переходов (см. рис. 6а). На рисунке 6б показано, что большинство переходов были измерены с отношением сигнал-шум, превышающим 10.

Рисунок 6в позволяет сделать заключение о количестве технических повторов в составе одного эксперимента. Видно, что в одном эксперименте в большинстве случаев исследовались 5 пептидов. Таким образом, в полученной базе данных имеется достаточный массив информации для исследования воспроизводимости регистрируемых масс-спектрометрическим детектором параметров в серии технических повторений.

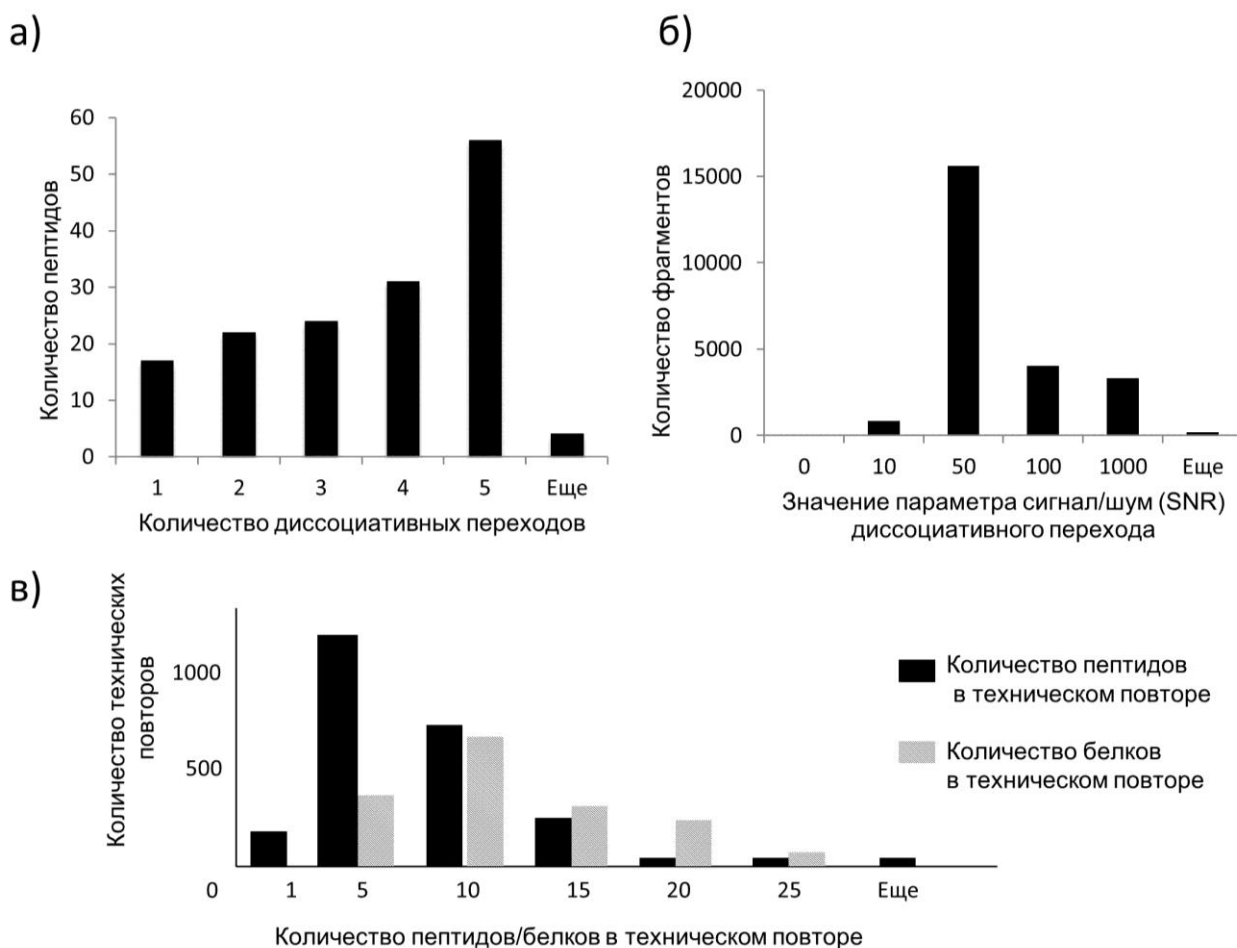


Рисунок 6. Гистограммы, характеризующие базу данных протеотипических пептидов: (а) распределение количества числа детектированных переходов на пептид; (б) распределение количества переходов с заданным значением параметра сигнал-шум; (в) распределение количества пептидов/белков в одном техническом повторе.

Доступ к базе данных был реализован в виде веб-ориентированного приложения, обеспечивающего интерактивную работу с результатами масс-спектрометрических экспериментов посредством интернет-обозревателя. Интерфейс доступа включал элементы, соответствующие объектам разработанной структуры данных.

Стартовая страница базы данных, доступная по адресу www.pikb18.ru, содержит статистические сведения об объеме внесенных данных. Пользователь может получить сведения о количестве экспериментов и технических повторов, а также о количестве детектированных белков и пептидов. Кроме того, на главной странице приведена аннотация использованного биологического образца

(количество, источник и пр.).

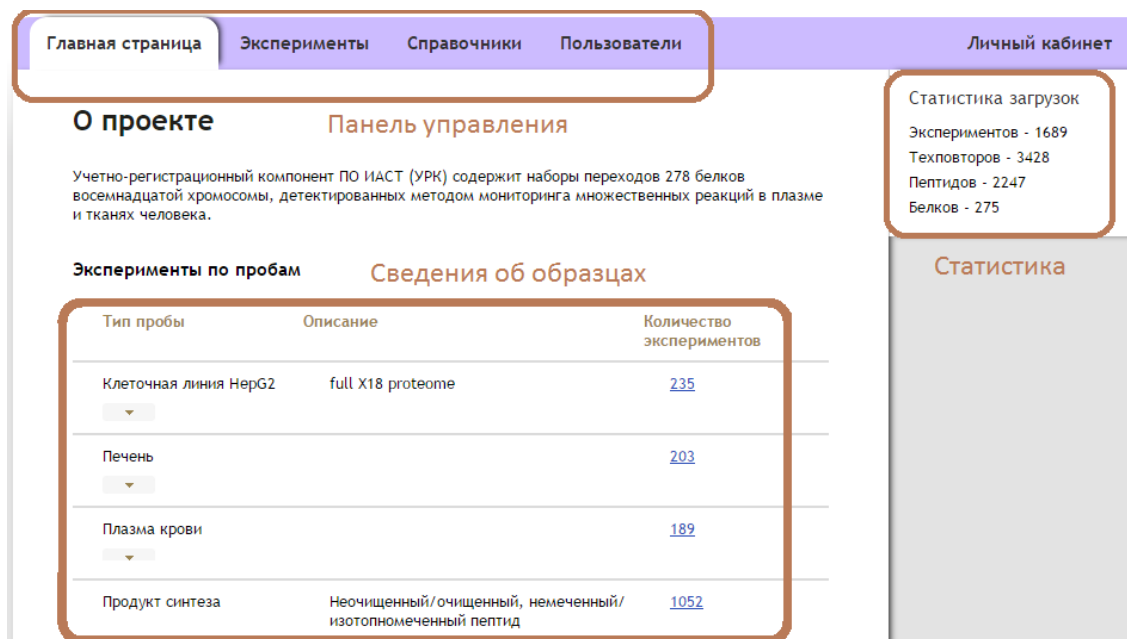


Рисунок 7. Стартовая страница разработанной базы данных.

В примере на рисунке 7 в разделе «статистика загрузок» показан один из промежуточных этапов пополнения базы данных протеотипических пептидов. Видно, что на 1689 экспериментов приходится более трех тысяч технических повторов, то есть примерно 2,1 технических повтора на один эксперимент. Количество пептидов превышает две тысячи. Это означает, что, в среднем, в одном эксперименте проводились измерения одного или двух пептидов (или одного пептида, но в разных зарядных состояниях).

Из статистики по образцам, приведенной на рисунке 7, следует, что большинство экспериментов (более 500) выполнено на растворах синтетических пептидов. Измерения, проводимые на синтетических пептидах, необходимы для создания библиотеки эталонных масс-спектрометрических сигналов. Однако в отличие от панорамных масс-спектрометрических методов эти библиотеки нельзя использовать для автоматической идентификации белков. Библиотеки используются при ручной обработке данных для сравнения полученных сигналов с эталонными. Также серии разведений пептидов в буферном растворе используются в структуре данных как составные элементы объекта «калибровочное уравнение». (см. рис. 5), с помощью которого производятся количественные измерения.

Обращение к экспериментальным данным, загруженным в систему, осуществляется с вкладки панели управления «Эксперименты». Если описанные выше пакеты подпрограмм предназначены для массовой обработки и интерпретации результатов большого количества экспериментов, то веб-ориентированная система предназначена для выполнения одиночных запросов и визуализации результатов их выполнения. По умолчанию отображается полный перечень экспериментов, содержащихся в базе данных. Для формирования запроса к базе данных использовали панель фильтров, расположенную в правой части экрана, а также поисковую строку. Панель фильтров позволяет отбирать эксперименты в соответствии с условиями их проведения (тип прибора, использованное для обработки ПО, тип биологического материала), а поисковая строка позволяет отбирать эксперименты в соответствии в кодом доступа UniProt AC белка и пептидной последовательностью.

Эксперименты (1689) - всего экспериментов в БД: 12(0) [Добавить](#)

по запросу: 12(0) Поисковый запрос: Q9P0L0

ID	Белки	Пептиды	Техновторы	Оператор	Проведен	Статус
239954 operational_testing_calibration_10_10-e9	1 (1)	1	2	Ильгисонис Екатерина	05.05.2013	опубликован 17.05.2013
Программное обеспечение: Mass Hunter Qualitative Analysis B 2.0. Инструмент: Agilent 6410 QQQ LC-ChipCube Положительный результат Пробы: Продукт синтеза: crude						
	Белок	Пептиды	m/z родительского иона	Заряд	Число транзичий (зип.)	
	Q9P0L0	HEQIVLDPPTLTK	540.29	3	5	
227323_Q9P0L0_operational_testing_cycle1	1 (1)	1	3	Ильгисонис Екатерина	16.04.2013	опубликован 22.04.2013
Программное обеспечение: Mass Hunter Qualitative Analysis B 2.0. Инструмент: Agilent 6410 QQQ LC-ChipCube Положительный результат Пробы: Клеточная линия HepG2: 25 million cells						
	Белок	Пептиды	m/z родительского иона	Заряд	Число транзичий (зип.)	
	Q9P0L0	VAHSKPGSTSTASFR	550.26	3	5	
Q9P0L0P1C_synthesis_25_E01_operational_testing_cycle2	1 (1)	1	3	Ильгисонис Екатерина	14.02.2013	опубликован 21.02.2013
Программное обеспечение: Mass Hunter Qualitative Analysis B 2.0. Инструмент: Agilent 6490 triple quadrupole lc/ms Положительный результат Пробы: Продукт синтеза: crude, 25 synthesis (A. Moisa)						
	Белок	Пептиды	m/z родительского иона	Заряд	Число транзичий (зип.)	
	Q9P0L0	VAHSKPGSTSTASFR	824.88	2	21	

Целевой белок, пептид и его m/z

Сведения о биологическом материале

Список экспериментов

Настройка фильтров

Дата проведения: не имеет значения [выбрать](#)

Оператор: не имеет значения [выбрать](#)

Режим сканирования: не имеет значения [выбрать](#)

Измерительный прибор: не имеет значения [выбрать](#)

Программное обеспечение: не имеет значения [выбрать](#)

Тип пробы: не имеет значения [выбрать](#)

Значение пробы: не имеет значения [выбрать](#)

Условие выбора эксперимента: не имеет значения [выбрать](#)

Статус эксперимента: не имеет значения [выбрать](#)

[Обновить](#) [Сбросить](#)

Рисунок 8. Формирование запроса к базе данных с использованием веб-ориентированного интерфейса. Настройка фильтров. Выборка экспериментов, соответствующая запросу по идентификатору белка «Q9P0L0».

Так, на рисунке 8 представлен результат выполнения запроса по белку «Q9P0L0». Запрос позволяет формировать выборку экспериментов, для каждого из которых содержится краткая характеристика: тип биологического материала, количество белков, детектированных в рамках эксперимента, перечень их пептидов. Для обращения к экспериментам на экране предоставляются гипертекстовые ссылки, обеспечивающие переход на страницу эксперимента. Используя заголовок таблицы, пользователь имеет возможность сортировать результирующий массив по названиям экспериментов, количеству детектированных белков, оператору, дате проведения эксперимента, а также статусу эксперимента. Система поддерживает опубликованные эксперименты, черновики, не содержащие экспериментальных данных, а также позволяет сохранять сведения об удаленных экспериментах. Данный функционал позволяет использовать систему при планировании новых экспериментов, а также для доступа к экспериментальным данным без использования специализированного ПО (Mass Hunter, Skyline и т.п.).

Если результатом запроса является один эксперимент, то система возвращает страницу, содержащую карту фрагментации, сведения о пробоподготовке, ссылки на технические повторы. Таким образом, выполнение запроса предоставляет пользователю информацию, достаточную для воспроизведения эксперимента.

В системе содержатся записи о каждом этапе масс-спектрометрического эксперимента. Процесс пробоподготовки отражен в виде протокола, загружаемого в виде отдельного файла свободного формата. Планирование эксперимента для исследования целевого белка по выбранному пептиду описывается с помощью карты фрагментации. При этом система сохраняет связь между планированием эксперимента и экспериментальной частью, отмечая на карте фрагментации фрагментные ионы, которые были детектированы в технических повторах (см. рис. 11). История редактирования записи в базе данных позволяет установить, когда именно был проведен эксперимент.

Главная страница Эксперименты Справочники Пользователи Личный кабинет

[К списку экспериментов](#) [Редактировать](#) [Удалить](#) [Удалить полностью](#) [Файлы](#) [Экспорт](#) [Скрыть](#)

156747_Q9P0L0_operational_testing_cycle1

Поместить наверх: ☒ Условия ☐ Результаты

Дата проведения: 01.02.2013
 Оператор: [Ильгисонис Екатерина](#)
 Статус: опубликован (07.03.2013)
 Результаты эксперимента: Успешней

Создан 7 марта 20
 Изменен 7 марта 20
 Автор последних изменений: джерчук Валентина

История редактирования записи в БД

Условия проведения эксперимента

Режим сканирования	статический MPM	Инструмент	Agilent 6410 QQQ LC-ChipCube
Белки	Q9P0L0	Программное обеспечение	Mass Hunter Qualitative Analysis B 2.0.
		Проба	клеточная линия hepg2: 25 million cells

Основная информация о пептидах

RT (theor.) - теоретическое время удержания, Количество фрагментов - теоретическое количество фрагментов для пептида в карте фрагментации, Количество транзиций - количество уникальных фрагментов, детектированных для пептида во всех технических повторах

Пептид	Белок	Код родительского иона	Mw (Da)	m/z	Заряд
VAHSDKPGSTSTASFR	Q9P0L0	Q9P0L0	1646.8	550.26	3

Целевой белок, пептид и его m/z

Карта фрагментации

Код родительского иона	Последовательность	m/z родительского иона	MS1 Res	Product ion m/z	Заряд Product ion	MS2 res	Collision energy	RT (theor.)	Тип иона
Q9P0L0	VAHSDKPGSTSTASFR	550.26	Unit	581.30	1	Wide	23.28	Y	
				668.34	1	Wide	23.28		
				769.38	1	Wide	23.28		
				913.44	1	Wide	23.28		
				1,010.49	1	Wide	23.28		

отмечены фрагментные ионы, детектированные в эксперименте

Рисунок 9. Страница эксперимента, в котором был детектирован белок «Q9P0L0». Результат запроса – раздел планирования экспериментов.

Раздел экспериментальной части в результате выполнения запроса позволяет пользователю производить простейшую интерпретацию результатов эксперимента, например, оценить воспроизводимость профиля транзиций в технических повторах, сравнивая m/z наиболее интенсивных фрагментов технических повторов. Например, на рисунке 10 видно, что в эксперименте в двух технических повторах наиболее интенсивным является фрагмент с m/z 668,34, тогда как в третьем техническом повторе наиболее интенсивным является фрагмент с m/z 1010,49. База данных предоставляет возможность интерпретации отдельных технических повторов посредством перехода на страницы с детализированной информацией о них по гипертекстовым ссылкам.

Технические повторы

Количество транзиций - количество уникальных фрагментов, детектированных в техническом повторе

ID техповтора (Эксп. ID_номер техповтора)	Дата проведения	Количество транзиций	Скачать
156747_Q9P0L0_operational_testing_cycle1_#1	21.01.2013 20:42	5	Excel compound report
156747_Q9P0L0_operational_testing_cycle1_#3	21.01.2013 22:15	5	Excel compound report
156747_Q9P0L0_operational_testing_cycle1_#2	21.01.2013 21:29	5	Excel compound report

**Переход на страницы
с результатами технических повторов**

**Экспорт результатов
технического повтора**

Код родительского иона	Белок	m/z	Заряд	RT (exp.)	Высота пика	Ширина пика	Площадь пика
Q9P0L0	Q9P0L0	550.26	3	12.746(0.11)	226.0(46.0)	0.17(0.04)	685.33(42.0)

Характеристики техповторов		m/z фрагментного иона с наибольшей интенсивностью пика				
Код пептидного иона	Белок	ID техповтора	m/z	Заряд	Высота пика	Сигнал/шум
Q9P0L0	Q9P0L0	156747_Q9P0L0_#2	668.34	1	122.44	33.5
		156747_Q9P0L0_#1	1010.49	1	132.2	33.3
		156747_Q9P0L0_#3	668.34	1	131.46	29.9

Рисунок 10. Страница эксперимента, в котором был детектирован белок «Q9P0L0». Результат запроса – раздел экспериментальной части.

Используя соответствующую гипертекстовую ссылку, пользователь может переключиться на страницу с детализированной информацией о каждом техническом повторе или загрузить таблицу с перечнем характеристик каждого детектированного фрагмента.

Технический повтор 156747_Q9P0L0_operational_testing_cycle1 #2

Пептидные ионы

Код пептидного иона	Белок	Mw (Da)	m/z	Заряд	RT (exp.)	Высота пика	Ширина пика	Площадь пика
1568 (Q9P0L0)	Q9P0L0	1646.8	550.26	3	12.708	214.0	0.17	664.0

Фрагменты

Фрагмент	m/z	Заряд	Тип иона	Сигнал/шум	Высота пика	Ширина пика	Площадь пика
Фрагмент 3346	581.3	1	Y	31.4	119.6	0.303	520.0
Фрагмент 3352	769.38	1	Y	28.0	52.27	0.29	278.0
Фрагмент 2853	913.44	1	Y	29.8	79.08	0.299	317.0
Фрагмент 3349	668.34	1	Y	33.5	122.44	0.299	788.0
Фрагмент 3355	1010.49	1	Y	33.4	115.64	0.356	831.0

Хроматограмма фрагмента 3346

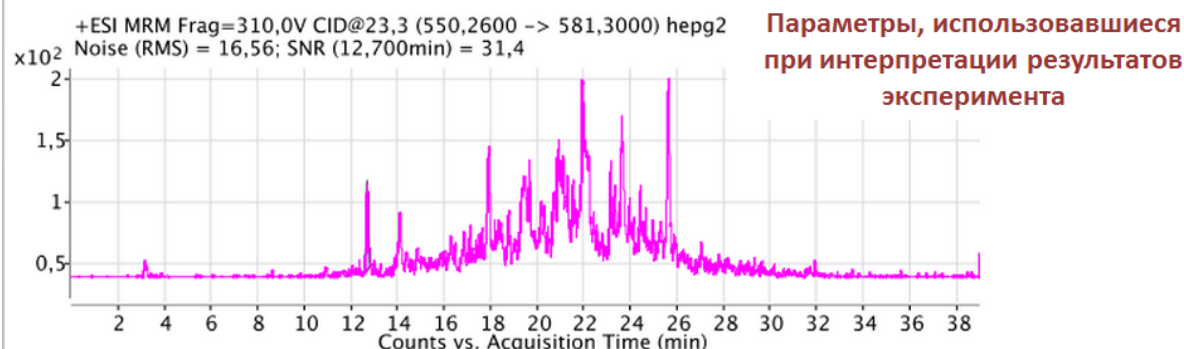


Рисунок 11. Технический повтор эксперимента, в котором был детектирован белок «Q9P0L0».

На рисунке 11 представлена детализированная информация о техническом повторе, в котором наиболее интенсивным фрагментом является фрагмент с m/z 668,34. Следующим по интенсивности являются фрагменты с m/z 581,30 и с m/z 1010,49 соответственно. Во втором техническом повторе данный профиль транзиций повторяется. Наибольший интерес представляет третий технический повтор, в котором профиль транзиций имеет следующий вид: 1010,49; 668,34; 581,30. Причем разница интенсивностей фрагментов с m/z 668,34 и с m/z 1010,49 составляет 25%, что позволяет судить о возникновении интерференции.

4.3. Оценка результатов детекции белков и пептидов в биологических образцах

В дальнейшей работе данные технических повторов использовали для оценки воспроизводимости результатов экспериментов. По определению,

воспроизводимость является характеристикой точности результатов измерений, определяемой близостью результатов повторных измерений. При этом полагали, что чем более воспроизводим результат в разных условиях постановки эксперимента, тем выше правдоподобие гипотезы о детектировании пептидов и, следовательно, тем точнее количественная оценка, выполненная методом мониторинга диссоциативных переходов. Изменчивость результатов измерений – мультифакторный признак, зависящий от методики эксперимента и свойств объекта измерений. Так, при исследовании протеомного состава биологического образца масс-спектрометрическим методом мониторинга диссоциативных переходов, разница в полученных результатах может быть обусловлена как действительными различиями между концентрациями белков, различием операторов или приборов, осуществляющих измерения, ошибками (погрешностями) измерений, биологической вариабельностью и сложностью образца.

Воспроизводимость является первостепенной характеристикой достоверных результатов. В данной работе анализировались измерения, имеющие минимальное количество методологических различий. Все результаты были получены с использованием одного прибора и обработаны одним оператором. Эксперименты были выполнены на стандартизированном образце исследований — клеточной линии HepG2. Оценка повторяемости измерений требует точного повторения всех условий их проведения. В случае же исследования биологических объектов невозможно однозначно определить всю совокупность параметров, влияющих на результаты эксперимента. Разработанная структура данных позволяет получить упрощенное представление о поведении объекта исследований в создаваемых условиях масс-спектрометрического измерения. Ключевые понятия разработанной структуры данных объективно описывают часть процессов и дают основу для последующей реконструкции детального описания.

Биологические объекты, в данном случае – протеом – чрезвычайно изменчивы, поэтому говорить об идентичности одного и того же объекта можно говорить лишь в малый промежуток времени. Практически реализованный

исследовательский процесс не позволяет оценить все параметры, так или иначе влияющих на результаты измерений (речь идет не об условиях выполнения эксперимента, но о параметрах, описывающих процессы в самом изучаемом объекте). Известно, что биологические эксперименты зачастую обладают неполной воспроизводимостью результатов измерений [Naegle, Gough, Yaffe, 2015]. Для уменьшения вариабельности результатов используют различные приемы: выполнение повторных экспериментов на одном и том же объекте, стандартизацию подготовки объектов к экспериментам, различного рода усреднения результатов и т.д. Количественное измерение содержания пептидов в биологическом материале подразумевает несколько технологических стадий, результаты выполнения каждой из которых имеют разную воспроизводимость. Таким образом, оцениваемая в данной работе воспроизводимость измерений в действительности является суммарной воспроизводимостью нескольких стадий исследования. Естественно, что наименее предсказуемой является воспроизводимость процессов, которые основаны на непосредственном взаимодействии с изменчивым объектом исследований – биологическим образцом: пробоподготовка, синтез и добавление стандартного раствора, а также хромато-масс-спектрометрические измерения. Таким образом, оценивая воспроизводимость результатов хромато-масс-спектрометрических измерений пептидов в данной работе, мы получаем представление о суммарной воспроизводимости всех стадий исследования.

В соответствии с требованиями FDA пороговым значением коэффициента вариации для исследуемых величин (время удержания, отношения сигнала к шуму, интенсивность) считали 20% [Carr и др., 2014]. Все измерения, коэффициент вариации которых превышали этот порог, при дальнейшем анализе не учитывали. Что касается коэффициента вариации профиля переходов, то данное понятие введено в данной работе по аналогии с коэффициентом вариации всех остальных величин. Однако, в соответствии с мировыми стандартами [Abbatiello и др., 2013], достоверными считались только те измерения, в которых профиль переходов пептидов полностью сохранялся во всех технических повторях. Остальные случаи

представляли интерес с точки зрения выдвижения и проверки гипотез о причинах изменения профиля.

В данной работе впервые проводится анализ массива данных, накопленного за значительный промежуток времени – более полутора лет, что позволяет оценить правдоподобие решений, принимаемых оператором масс-спектрометра. Это отличает разработанный подход от аналогов, реализованных в программных пакетах MProphet и MRMer, где предметом статистического анализа являются эксперименты, снятые за короткий промежуток времени практически в идентичных условиях, однако при этом заложенные в них обучаемые статистические модели до сих пор не получили широкого распространения. Поскольку биоматериал, инструмент, на котором проводились измерения, а также оператор оставались неизменными на протяжении полутора лет, можно говорить о том, что причиной выявленных противоречий в данных являются именно технические ограничения метода, а также сложность биологического образца, а не различия в использованных экспериментальных протоколах.

Были выявлены типы распределений значений коэффициента вариации для различных параметров, регистрируемых в ходе эксперимента, выполняемого методом мониторинга диссоциативных переходов. Для ряда параметров (например, время удержания на хроматографической колонке), как показано на рисунке 12а, подавляющее большинство значений имеют крайне низкий коэффициент вариации (менее 1%). Это объясняется тем, что в используемом для обработки первичных данных ПО Mass Hunter заложена функция определения по совпадению времен удержания группы пиков, соответствующих одному соединению. Это следует из процесса сопряжения хроматографии и масс-спектрометрической детекции, то есть если фрагменты действительно принадлежат одному пептиду, то они должны регистрироваться в одно и то же время. Другая группа параметров (см. рис. 12б и рис. 12в) характеризуется другим типом распределения коэффициента вариации, имеющим выраженный максимум. Например, коэффициент вариации интенсивности в наибольшем количестве экспериментов составляет от 50 до 150 процентов (см. рис. 12в). Разница в количестве экспериментов, используемом при

построении распределения, объясняется последовательным использованием фильтров и использованием на каждом шаге выборки экспериментов, прошедших предыдущие каскадные фильтры. Для проверки значения каскадных фильтров, использованных для выявления протеотипических пептидов с наименьшей погрешностью в количественной оценке, проводили сравнение результатов количественных измерений двух парных пептидов, принадлежащих к одному белку. Если они относятся к одному белку, то должны находиться в эквимольных концентрациях, то есть их концентрация в биоматериале не должна различаться существенным образом.

С точки зрения совпадения результатов количественных измерений парных пептидов, первоначальный массив данных были разбит на 4 группы пептидов, охарактеризованные в таблице 5.

Таблица 5. Результаты сегментации исходного массива данных.

Сегмент, №	Количество белков	Количество пептидов	Различие результатов количественных измерений пептидов
№1	156	312	<100%
№2	37	74	1 порядок
№3	24	48	2 порядка
№4	51	102	3 и более порядков

Из таблицы видно, что в 56% случаев результаты, полученные по двум пептидам одного и того же белка, не противоречат друг другу. Сочетание сравнения результатов количественных измерений двух парных пептидов и оценки воспроизводимости результатов масс-спектрометрической детекции с высокой вероятностью позволяет исключить ложноположительную детекцию протеотипических пептидов.

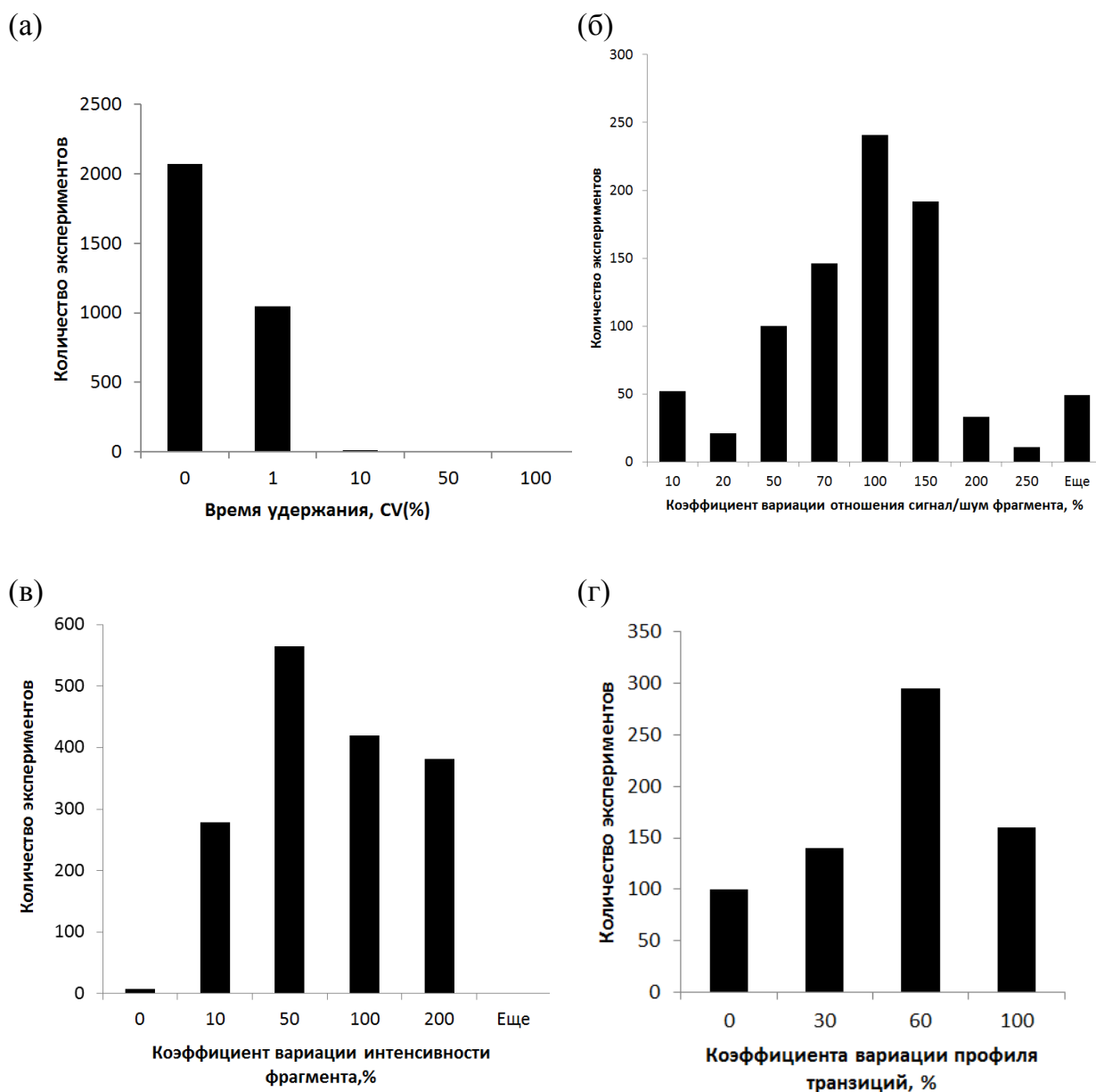


Рисунок 12. Гистограмма распределения коэффициента вариации в технических повторях (а) усредненного для всех фрагментных ионов пептида времени удержания; (б) отношения сигнал/шум фрагментных ионов; (в) интенсивностей фрагментных ионов пептидов; (г) профилей фрагментных ионов пептидов.

Исходя из вышеизложенного, был разработан каскад фильтров, в который вошли фильтры, связанные с воспроизводимостью интенсивности и профиля переходов, а также непротиворечивость измерений уникальных парных пептидов, относящихся к одному белку. Из исходного массива данных, который изначально насчитывал 2247 пептидов для 275 белков, в результате применения фильтра по интенсивности более полутора тысяч пептидов было отвергнуто. Применение

следующего фильтра, основанного на сопоставлении профилей переходов, сократила результирующую выборку до 86 пептидов (65 белков). Среди оставшихся белков только для 23 имелись данные о результатах измерений двух пептидов, относящихся к одному белку. Сегментация полученного массива позволила разделить его на четыре группы. В первую группу вошли 12 белков, для которых результаты измерений парных пептидов различались не более, чем в два раза. Вторая группа включала четыре белка с более существенными отличиями - в пределах одного порядка. В третьей группе допускались различия в пределах трех порядков, естественно за исключением тех случаев, которые вошли в состав первых групп. Эта группа состояла из шести пептидов трех белков. Наконец, в группу с чрезвычайно существенными различиями, превышающими три порядка величины, были отнесены четыре белка и восемь пептидов.

На рисунке 13 показана последовательность действий для анализа массива экспериментов, полученных при масс-спектрометрическом исследовании белков, кодируемых генами хромосомы 18. Согласно первой задаче, была разработана структура данных, предоставляющая возможность хранить и обрабатывать эксперименты, выполненные методом мониторинга диссоциативных переходов. Далее разработанная структура данных была реализована в виде информационного ресурса, включающего базу данных и веб-интерфейс для работы с ней. Посредством системы запросов к информационному ресурсу были выявлены протеотипические пептиды, пригодные для проведения количественных измерений белков, кодируемых генами хромосомы 18. Они вошли в состав результирующей выборки из 23 белков.



Рисунок 13. Каскад фильтров позволяет оценить воспроизводимость в технических повторях, сегментация – распределить данные по группам с разной степенью противоречивости результатов измерения протеотипических пептидов, относящихся к одному белку. 1- 4 – сегменты данных; ф1 – фильтр по отношению сигнал/шум; ф2 – по интенсивности; ф3 – по профилю переходов.

Представленный на рисунке 13 алгоритм был использован для формирования выборки пептидов белков, кодируемых генами хромосомы 18, пригодных для количественных измерений. Эксперименты, соответствующие этим пептидам, считали достоверными и правдоподобными, и поэтому пригодными для проверки молекулярно-биологических гипотез.

4.4. Анализ выборки протеотипических пептидов и соответствующих им белков

4.4.1. Репрезентативность выборки протеотипических пептидов и соответствующих им белков

Выборку, в которую вошли 23 белка, охарактеризовали с использованием широко применяемых в протеомике информационных ресурсов. Из БД UniProt использовали информацию об экспериментах, подтверждающих экспрессию генов и трансляцию мРНК. Из онтологии генов GO получали сведения об участии белков результирующей выборки в молекулярных процессах. С использованием этих ресурсов было показано, что полученная выборка репрезентативна для хромосомы

18. В таблице 5 приведены пептиды и белки, вошедшие в состав результирующей выборки, термины GO, которыми проаннотированы кодирующие их гены, а также сведения о количественном содержании в клеточной линии HepG2.

Таблица 6. Примеры белков и пептидов, вошедших в состав контрольной выборки (полный список белков и пептидов контрольной выборки в приложении).

UniProt AC (Название белка)	Пептидная последовательность	Количество копий пептида в клетке линии HepG2	Клеточная локализация, GO	Молекулярно- биологическая функция, GO
A6NNZ2 (Тубулин бета 8 подобный белок)	HGCYLTVAAlFR	2.52*10 ⁸	Цитоплазма [GO:0005737]	Активность ГТФаз [GO:0003924]
	LAVNMVPFPR	2.52*10 ⁸		
O60543 (Активатор клеточной смерти CIDEA)	GVMASSLQELISK	3.53*10 ⁵	Цитоплазма [GO:0005737]	Полиаденили- рование [GO:0044822]
	VTFDLYR	4.56*10 ⁵		
P18621 (60S рибосомальный белок L17)	YSLDPENPTK	4.34*10 ⁶	Цитоплазма [GO:0005737]	Активность кортикотропиново го рецептора [GO:0004978]
	EQIVPKPEEEVAQK	5.10*10 ⁶		
P48594 (Серпин B4)	LVLVNAIYFK	1.47*10 ⁶	Межклеточный контакт [GO:0005913]	Связывание ионов кальция [GO:0005509]
	VLHFDQVTENTTEK	1.47*10 ⁶		
Q01718 (Рецептор адренокортикотроп ного гормона)	NNSDCPR	1.93*10 ⁵	Внеклеточная локализация [GO:0005576]	Структурная активность молекул [GO:0005198]
	NMGYLKPR	1.37*10 ⁵		
Q02413 (Десмоглеин-1)	YQGTLISIDNLR	7.11*10 ²	Цитоплазма [GO:0005737]	-
	LADISLGK	8.89*10 ²		
Q02487 (Десмоглеин-2)	TNEGVLCVVKPLNYEEK	6.15*10 ³	-	Связывание ионов кальция [GO:0005509]
	SFTILLSNTENQEK	1.30*10 ⁵		
Q15846 (Кластерин - подобный белок-1)	APDHGGLISK	1.41*10 ³	Ядро [GO:0005730]	Связывание повреждённой ДНК [GO:0003684]
	IYQFLFPFHEDNEK	6.90*10 ³		
Q6UXH8 (Коллаген - и кальций - связывающий белок)	GSMGPMGSPDLSHIK	2.45*10 ⁴	Цитоплазма [GO:0005737]	Полиаденили- рование [GO:0044822]
	YPNDTGHEK	1.37*10 ⁴		
Q9Y2V3 (Белок ретикулярного гомеобокса)	DDGILGTFFAER	9.68*10 ³	Ядро [GO:0005634]	Связывание повреждённой ДНК [GO:0003684]
	FPLDEADPR	1.48*10 ⁵		

Обращаясь к таблице 6, прежде всего отметим, что каждому белку соответствует два целевых пептида. Данное условие было задано при выполнении этапов фильтрации информационного массива. Количество аминокислотных остатков в пептиде колеблется от 7 до 20, со средним значением 12. Таким образом, отношение массы к заряду для отобранных пептидов оптимальным образом отвечают техническим возможностям тройного квадрупольного детектора. Следует отметить, что в составе представленных в таблице 5 пептидов отсутствуют так называемые «недорезы» - продукты частичного гидролитического расщепления. Это связано с использовавшимся способом отбора протеотипических пептидов [Zgoda и др., 2013]. С использованием программы был проведен расчет теоретического времени удержания пептида на хроматографической колонке при условиях, указанных в таблице 2. Расчет показал, что пептиды равномерно сходят в диапазоне от первой до 38-ой минуты, наибольшее количество пептидов сходит в интервале от 13-ой до 18-ой минут.

Указанные в таблице 6 пептиды часто подвергают идентификации методами панорамной масс-спектрометрии. Для оценки их изученности был использован ресурс PRIDE [Jones и др., 2006], в котором в мае 2015 года было депонировано более 50 тысяч спектров. Итак, для пептида LAVNMVPFPR в PRIDE было найдено наибольшее количество экспериментов. В среднем пептиды были идентифицированы в 13 экспериментах. Пептиды YPNDTGHEK, APDHGGLISK, IYQFLFPFHEDNEK, FPLDEADPR, NMGYLKPR, GVMASSLQELISK не были детектированы ни в одном панорамном масс-спектрометрическом эксперименте.

Рассмотренные данные о физико-химических свойствах пептидов, об их встречаемости в репозиториях масс-спектрометрических данных позволяют сделать следующий вывод. Подобранные фильтры на параметры технологических операций, в основе которого лежит метод МДП, корректно отражают особенности хроматографического разделения, ионизации и диссоциации протеотипических пептидов. То есть, нет оснований полагать, что на отбор белков каким-либо образом повлияли физико-химические свойства пептидов, кроме тех, которые определяли его способность к детекции масс-спектрометрическими методами.

Приведенные в таблице 6 пептиды достаточно разнообразны по молекулярному весу, аминокислотному составу, расчетному времени удержания и изоэлектрической точке. Что касается белков, то характеризующие их термины онтологии генов не имеют ярковыраженных отличий по сравнению со всем протеомом. Таким образом, можно предполагать, что полученная в результате использования системы фильтров выборка репрезентативно отражает как свойства хромосомы 18, так и всего протеома человека.

На рисунке 14 показано, что подразумевается под репрезентативностью выборки белков, обеспечиваемой хромосомотцентричным подходом. В хромосоме 18 доля транслируемых в белки генов превышает 65%, причем в других хромосомах данный показатель находится примерно на таком же уровне [Lane и др., 2014]. Как следует из рисунка 14а, соотношение в полученной нами выборке остается неизменным: отобранные с использованием каскада фильтров белки преобладают в категории «протеомный уровень». Аналогичная ситуация наблюдается на рисунке 14б, где приведены данные по количеству копий пептидов в единичной клетке HepG2. Видно, что в полученной выборке, как и во всей хромосоме 18, максимальное количество пептидов представлено в количестве 100 тысяч копий на клетку HepG2.

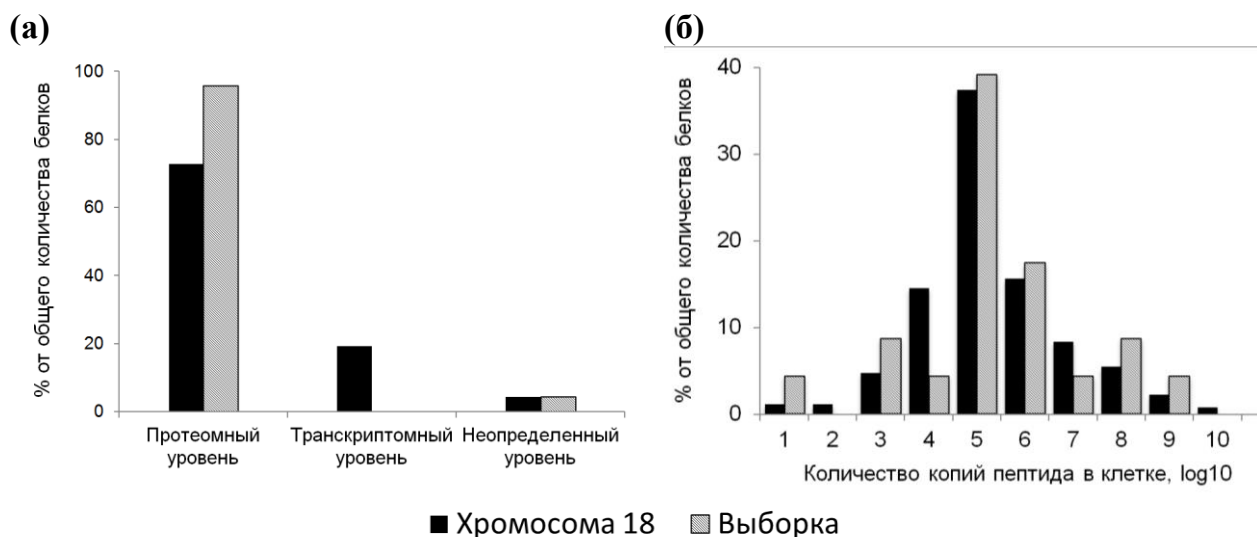


Рисунок 14. Сопоставление хромосомы 18 (276 белков) и выборки (23 белка):
 (а) уровень реализации генетической информации согласно ресурсу UniProt;
 (б) распределение количества копий пептидов в расчете на единичную клетку HepG2.
 Выборка – набор белков, полученный в результате анализа свойств протеотипических пептидов (см. табл. 5)

Особый интерес представляет белок Q8IYS8, для которого было с использованием разработанного алгоритма впервые была показана возможность детектировать его в клеточной линии HepG2 на протеомном уровне с высокой воспроизводимостью. Результаты детекции этого белка в клеточной линии были подтверждены с использованием изотопно-меченого пептида. Хроматографические пики, полученные для эндогенного и изотопно-меченого пептидов в клеточной линии HepG2, представлены на рисунке 15. Предположительно, белок Q8IYS8 локализован в клеточном центре и принимает участие в формировании веретена деления клетки. Ранее существование белка Q8IYS8 не было доказано экспериментальными методами.

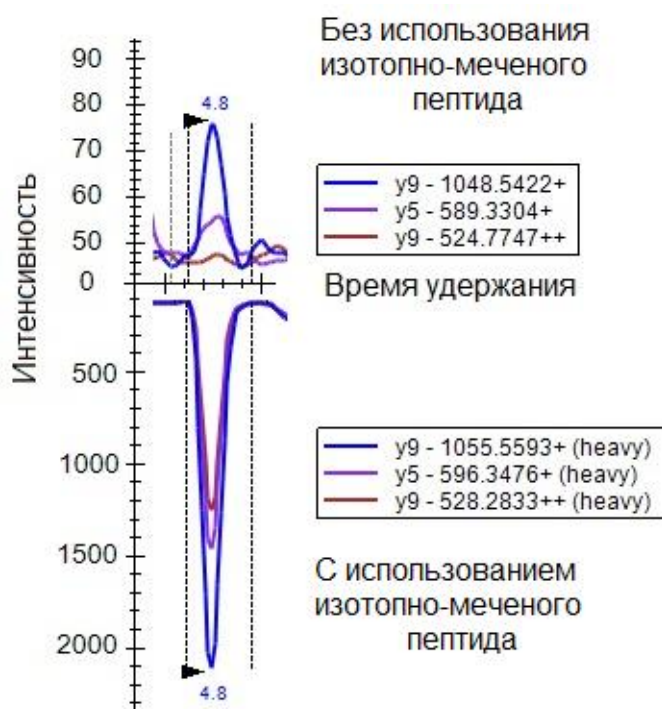


Рисунок 15. Хроматограмма белка Q8IYS8, для которого было доказано его существование на протеомном уровне в клеточной линии HepG2 с использованием изотопно-меченого пептида и без него.

Сопоставляя аннотации GO кодируемых хромосомой 18 генов с аннотацией всего генома с использованием инструмента GOrilla, с вероятностью ошибки менее 10^{-4} получили, что гены хромосомы 18 обеспечивают формирование межклеточных десмосомальных контактов. Как минимум продукты 19 генов (7% от общего числа генов хромосомы) участвуют в образовании межклеточных контактов. Для выборки из 23 белков, сформированной на основе оценки правдоподобия масс-спектрометрической детекции, получено, что 13% генов принадлежат к категории «межклеточный» контакт. Это означает, что в плане клеточных функций контрольная выборка *конгруентна* исследуемой хромосоме.

Результаты применения каскадных фильтров позволяют получать несмещенную выборку как в отношении сведений об экспрессии генов, так и в отношении количества копий белковых продуктов в единичной клетке (см. рис. 14). С учетом этого провели сопоставление выборки с масс-спектрометрическими данными. Использовали данные, полученные направленным методом в плазме крови человека [Ponomarenko и др., 2014; Zgoda и др., 2013] и панорамным методом в клеточной линии HepG2 [Schaab и др., 2012]. При сопоставлении

данных по плазме крови получили, что 15 из 23 белков в составе результирующей выборки детектируются как в клеточной линии HepG2, так и в плазме крови (при этом их концентрации не совпадают). Отсутствие корреляции наблюдалось также при сопоставлении с результатами экспериментов, опубликованных в базе данных MaxQB [Schaab и др., 2012] для всей хромосомы 18 (см. рис. 15а). Корреляция повышалась при сопоставлении в пределах контрольной выборки ($R^2 = 0,59$) (рис. 16б), однако, наблюдаемая на рисунке 16б картина показывает, что достоверность корреляционной связи низкая, так как основное количество точек сгруппировано в середине диаграммы.

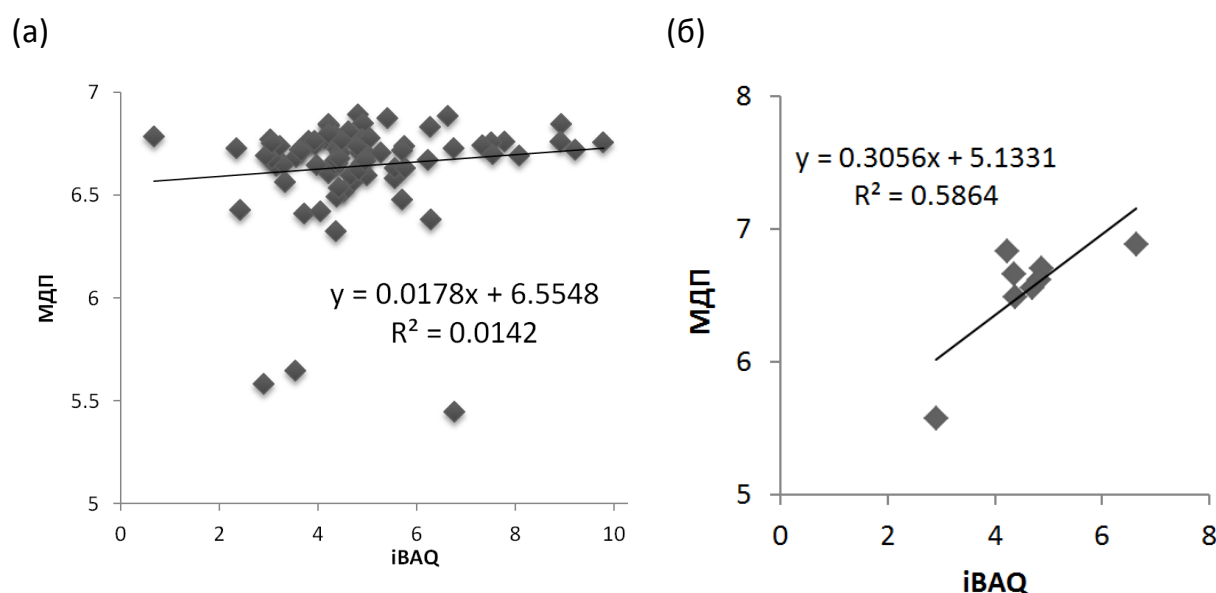


Рисунок 16. Корреляция результатов измерения количественного содержания белков в клетках HepG2, полученных направленным методом мониторинга диссоциативных переходов и панорамным масс-спектрометрическим методом iBAQ: (а) для 80 продуктов генов хромосомы 18; (б) для выборки из 23 белков, полученной в результате отбора протеотипических пептидов.

Низкий уровень корреляции, вероятно, связан с техническими различиями методов направленной и панорамной масс-спектрометрии и использованием разных пептидов при детектировании одних и тех же белков. Например, для белков, представленных на рисунке 16б, только 7 пептидов, использованных для измерения количественного содержания, совпадают. Белок P18621 был детектирован методом МДП с использованием пептидов EQIVPKPEEEVAQK и

YSLDPENPTK, тогда как при использовании метода iBAQ детектировали пептиды EQIVPKPEEEVAQK, GIDVDSIVIEHIQVVK, QWGWWTQGR, SAEFIINMIK.

Дальнейший анализ выборки белков, приведенной в таблице 6, проводили с использованием ресурса Plasma Proteome DB [Nanjappa и др., 2014]. В этом ресурсе обнаружили сведения только для трех белков хромосомы 18. Концентрации белков десмоглеина-1 (Q02413), десмоколлина-2 (Q02487) и серпина В4 (P48594) составили 14 нг/мл, 2,7 нг/мл и 2 нг/мл, соответственно. С результатами измерений, размещенными в разработанной базе данных протеотипических пептидов, указанные значения совпадали с ошибкой в пределах 50%, что для количественных масс-спектрометрических экспериментов является вполне допустимым [Carr и др., 2014].

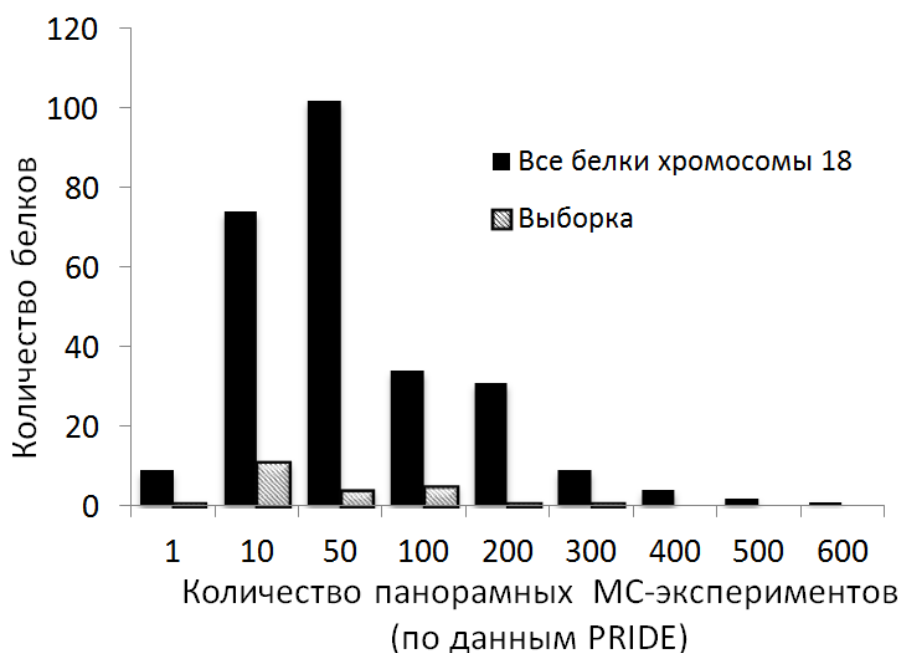


Рисунок 17. Количество идентификаций в PRIDE белков 18 хромосомы и результирующей выборки.

При этом все белки контрольной выборки были детектированы хотя бы в одном панорамном масс-спектрометрическом эксперименте, загруженном в репозиторий PRIDE. В среднем белки контрольной выборки были детектированы в 50 панорамных экспериментах (см. рис. 17), что свидетельствует о том, что данные

белки могут использоваться для масс-спектрометрических исследований, в том числе для количественных измерений.

4.4.2. Интерференция протеотипических пептидов

Принятой практикой оценки детекции пептида является сравнение профилей переходов эндогенного пептида, измеренного в биологическом материале, и синтетического пептида, измеренного в буферном растворе [Gallien и др., 2012]. Сопоставление профилей всех пептидов белков хромосомы 18, измеренных в буферном растворе и биологическом материале показало, что только 20 пептидов удовлетворяют этому требованию. Очевидно, что причина несовпадения профилей заключается в наличии смеси химических соединений, называемой матрицей, в биологическом образце. За счет наличия интерферирующих соединений соотношение регистрируемых интенсивностей меняется, то есть детектируемый масс-спектрометром сигнал представляет собой сумму сигналов множества соединений, совпадающих по своим масс-зарядным характеристикам с искомым соединением (пептидом) [Бао и др., 2013]. Источником сигнала могут быть пептиды и их фрагменты, низкомолекулярные соединения. Изменчивость суммарного многокомпонентного сигнала зависит от количества компонентов и их стабильности. Наиболее изменчивыми компонентами сигнала от пептида являются сигналы от других пептидов в силу многообразия протеоформ в составе протеома [Smith, Kelleher, Proteomics, 2013]. Явление интерференции объясняет различие между профилями транзиций пептида, полученных в различных типах биологического материала. Поэтому использование унифицированных конфигурационных методов для детекции одного пептида в разных биологических материалах может привести к получению недостоверных результатов. Разница в составе фоновой биологической матрицы может влиять на состав возможных интерферирующих компонентов. Следовательно, отбор пептидных и фрагментных ионов для конфигурационного метода, загружаемого в масс-спектрометр, должен производиться с учетом специфики биоматериала.

Для изучения влияния явления интерференции на проведение количественной оценки пептидов результирующей выборки использовали ПО

SRM Collider. SRM Collider является инструментом, позволяющим оценить вероятность возникновения интерференции. Пользователь производит оценку самостоятельно на основании визуализации сведений о количестве потенциальных интерферирующих пептидных и фрагментных ионов. В используемой статистической модели детекция всех пептидов считается равновероятной. Это противоречит существующим представлениям о протеоме человека, определяющим протеом как совокупность белков с концентрациями, варьирующимися от 10^{-18} до 10^{-6} [Archakov и др., 2011]. Пептиды имеют разную концентрацию в различных типах биологического материала (что объяснимо с точки зрения биологии), и, следовательно, имеют разную вероятность быть детектированными. При поиске веществ в малой концентрации естественно, что вероятность обнаружения зависит от концентрации и чувствительности прибора [Archakov и др., 2012]. Следовательно, вероятностная модель, использованная в SRMCollider, не может быть использована при планировании эксперимента. Однако при пост-экспериментальной обработке данное программное обеспечение может использоваться для проверки гипотезы о возникновении интерференции, что и было сделано в данной работе.

Обратим внимание на еще один важный аспект приводимых в таблице 6 данных. Видно, что для ряда белков количественные данные, полученные с использованием разных протеотипических пептидов, существенно отличаются. Разница между измеряемыми концентрациями парных пептидов может объясняться двумя причинами. Во-первых, возникновением явления интерференции в случае хотя бы одного из двух пептидов, то есть наличием соединений, имеющих масс-зарядные характеристики, совпадающие с масс-зарядными характеристиками целевого пептида. Во-вторых, неуникальностью целевого пептида в рамках протеома исследуемого организма. В следующем разделе будет рассмотрен вопрос об использовании количественных данных для выявления сплайс-опосредованных вариантов белков.

Для выявления характеристик, отличающих пептиды из выборки от прочего массива пептидов, был проведен анализ оценки вероятности возникновения

интерференции. В работе [Mallick и др., 2007] было показано, что массив триптических пептидов человека является однородным с точки зрения физико-химических свойств. Достоверных связей между аминокислотным составом пептида и вероятностью его детектирования в масс-спектрометрическом эксперименте выявлено не было. С точки зрения возможных масс-зарядных характеристик и потенциального времени удержания на хроматографической колонке множество пептидов является дискретным, что объясняется конечным набором атомов с определенными массами, входящими в состав пептидов. Это следует из анализа распределения потенциальных интерферирующих пептидных ионов в координатах масс-зарядная характеристика/расчетное время удержания (см. рис. 17). Видно, что точки упорядочены вдоль вертикалей, соответствующих масс-зарядным характеристикам. Данный рисунок позволяет оценить сложность использования так называемой точной массово-временной метки в протеомных исследованиях. Одной масс-зарядной характеристике и одному времени удержания может соответствовать несколько пептидов. В данной работе SRM Collider использовали для оценки вероятности возникновения различия между парными пептидами вследствие интерференции. Анализировали вероятность возникновения интерференции для 46 пептидов из результирующей выборки, а также для случайно выбранных 43 пептидов, не вошедших в ее состав из-за несоответствия системе разработанных фильтров.

Детектируемый в отдельно взятый момент времени масс-спектрометром сигнал представляет собой суммарный сигнал от множества соединений, имеющих масс-зарядную характеристику, на регистрацию которой настроен прибор и физико-химические свойства, обеспечивающие соответствующее время удержания. Источником сигнала могут быть как белковые продукты и их фрагменты, так и низкомолекулярные соединения. Количество пептидных ионов или их фрагментов, которые потенциально могут интерферировать с сигналом от целевого пептида, зависит от нескольких параметров. При этом стоит отметить, что пептиды меньшей длины, характеризуются наличием большего количества интерферирующих фрагментов. Это объяснимо с точки зрения комбинаторики, поскольку чем меньше

длина пептида, тем меньше вероятность того, что он и его фрагменты будут уникальными в рамках протеома одного организма.

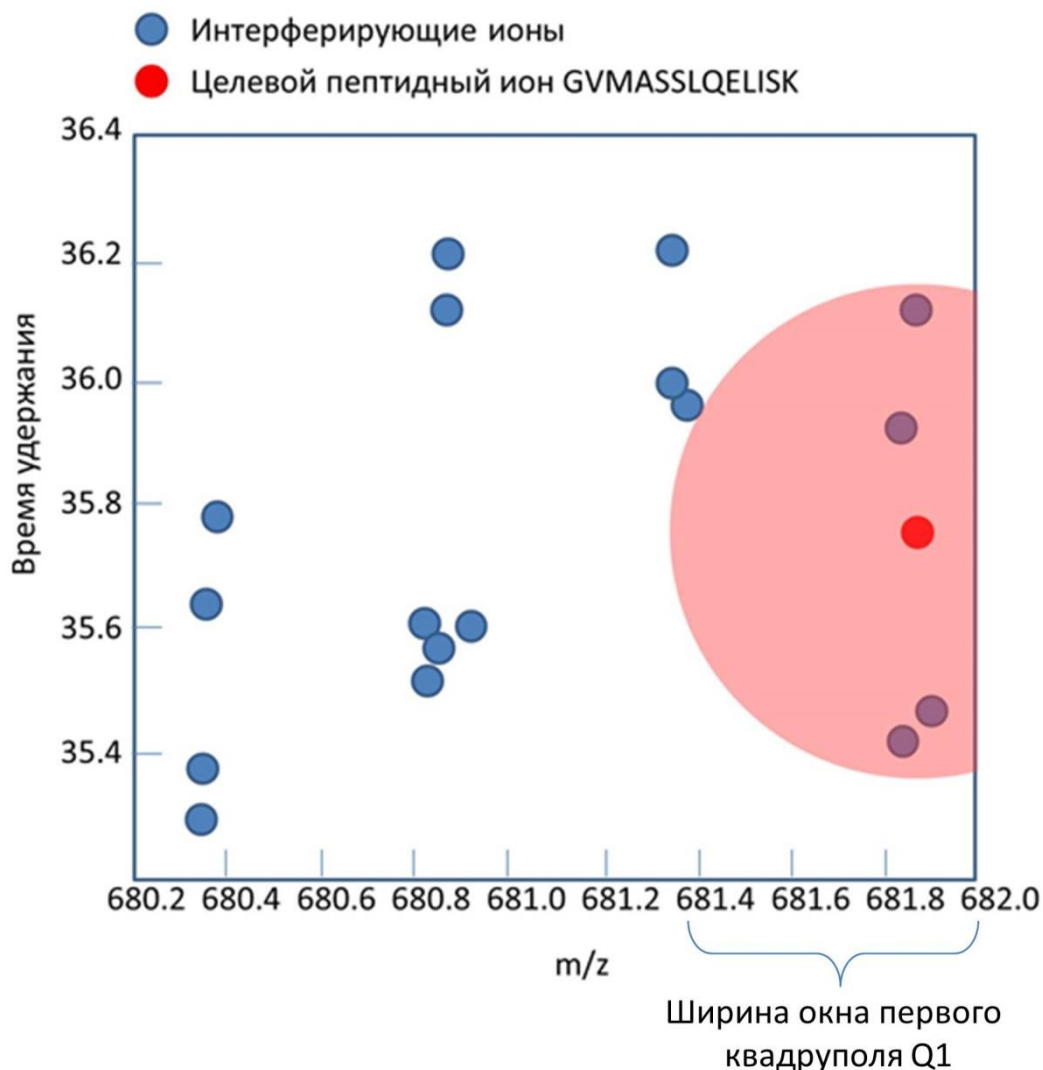


Рисунок 18. Диаграмма ионов, теоретически способных интерферировать с ионом целевого пептида GVMASSLQELISK. Диаграмма построена с использованием ПО SRM Collider.

От точности масс-спектрометрической детекции масс-зарядных характеристик зависит количество потенциальных интерферирующих фрагментов. На графике (см. рис. 18) показано, как меняется количество потенциально интерферирующих пептидных ионов в зависимости от ширины окна первого квадруполь Q1 для пептидов экспериментальной выборки (рис. 19а) и случайного набора (рис. 19б). Чем шире окно квадруполь, тем ниже специфичность проводимых измерений. Видно, что чем меньше окно первого квадруполь (0,01-0,02), тем меньше количество потенциальных интерферирующих фрагментов.

Разрешающая способность масс-спектрометров не позволяет устанавливать такие низкие значения ширины окон; обычно используют окна квадруполь Q1 и Q3 в диапазоне 0,7–1,0. При такой ширине большинство пептидных ионов находится в области, где количество интерферирующих фрагментов максимально (см. рис. 17).

Ширина окна является основным фактором, влияющим на интерференцию [Domon, 2012]. Нормальной практикой считается использование ширины окна первого и третьего квадруполь 0,7. В работах [Gallien и др., 2012; Teleman и др., 2012] было показано, что такое окно достаточно для детекции 158 пептидов дрожжей и 163 пептидов *S. pyogenes strain*. С другой стороны, такие масштабы на порядок меньше, чем используемые в настоящей работе данные о 276 белках и >1000 пептидах, полученных хромосомотричным подходом [Ponomarenko и др., 2014; Zgoda и др., 2013].

Специфичность масс-спектрометрического исследования зависит от разрешающей способности детектора, то есть способности различать две близкие массы при установленных в эксперименте окнах Q1 и Q3. Исследование количественного содержания биомаркеров требует более точных измерений, однако, в вышеуказанных работах ширина окна квадруполь используется как при поиске, так и при валидации биомаркеров. В работе [Kim и др., 2013] ширина 0,7 использовалась при анализе 28 кандидатных биомаркеров гепатоцеллюлярной карциномы. Sajic и соавторы использовали окно 0,7 при исследовании 95 потенциальных маркеров немелкоклеточного рака легкого [Sajic, Liu, Aebersold, 2014]. Меньшее значение (0,35) использовалось в работе [Kim и др., 2015] при детекции 54 низкокопийных маркеров заболеваний печени в плазме крови человека.

В данной работе использовался массив данных, полученных при установленной ширине окна первого и третьего квадруполь 0,5 [Ponomarenko и др., 2014; Zgoda и др., 2013]. С одной стороны, это совпадает с описанной выше мировой практикой, с другой стороны, свидетельствует о вероятности возникновения интерференции при детекции пептидов. При исследовании влияния интерференции на результаты измерений расчет интерференции осуществляли с

использованием ПО SRMCollider. При анализе данных, полученных с помощью ПО SRM Collider в применении к совокупности отобранных пептидов, изучался вопрос влияния интерференции на воспроизводимость результатов в технических повторях. Предположим, что в составе хроматографической фракции присутствует несколько интерферирующих соединений (как на уровне пептидного иона, так и на уровне фрагментов). Тогда интерференционная картина должна быть воспроизводима в технических повторях. Следовательно, такие параметры как время выхода фракции с колонки и относительная интенсивность фрагментных ионов не должны меняться, несмотря на явление интерференции, то есть, что картина интерференции должна повторяться в технических повторениях в неизменном виде.

Изложенная в работе [Sherman и др., 2009] концепция уникальной ионной сигнатуры может рассматриваться как аргумент в пользу более сложных стохастических процессов, в отличие от базовых представлений, полученных в результате масс-спектрического анализа чистых растворов веществ. Стохастический компонент связан с процессом хроматографии, в ходе которого свойства разделяющего соединения–носителя могут изменяться при повторении эксперимента [Gorshkov и др., 2006]. Состав компонентов фракции изменяется, поэтому в каждом техническом повторе они дают различную интерферирующую картину. При этом выделить вклад отдельных интерферирующих фрагментов, как ранее мы показали на рисунке 18, не представляется возможным.

Стохастический характер масс-спектрометрического эксперимента проявляется особенно ярко, если рассматривать идентичные по условиям выполнения эксперименты, выполненные одним оператором с использованием одного прибора, но в разное время.

Для выявления свойств, характерных для пептидов контрольной выборки, был проведен сравнительный анализ со случайным набором пептидов. В качестве случайного набора данных использовали пептиды, соответствовавшие разработанным фильтрам (воспроизводимость интенсивности, профиля транзиций и времени удержания в технических повторях), парный пептид которых не прошел

систему фильтров. Таким образом, для случайного набора данных было отобрано 20 пар пептидов. Половина из отобранных пептидов соответствовала разработанному фильтру, а другая половина – нет.

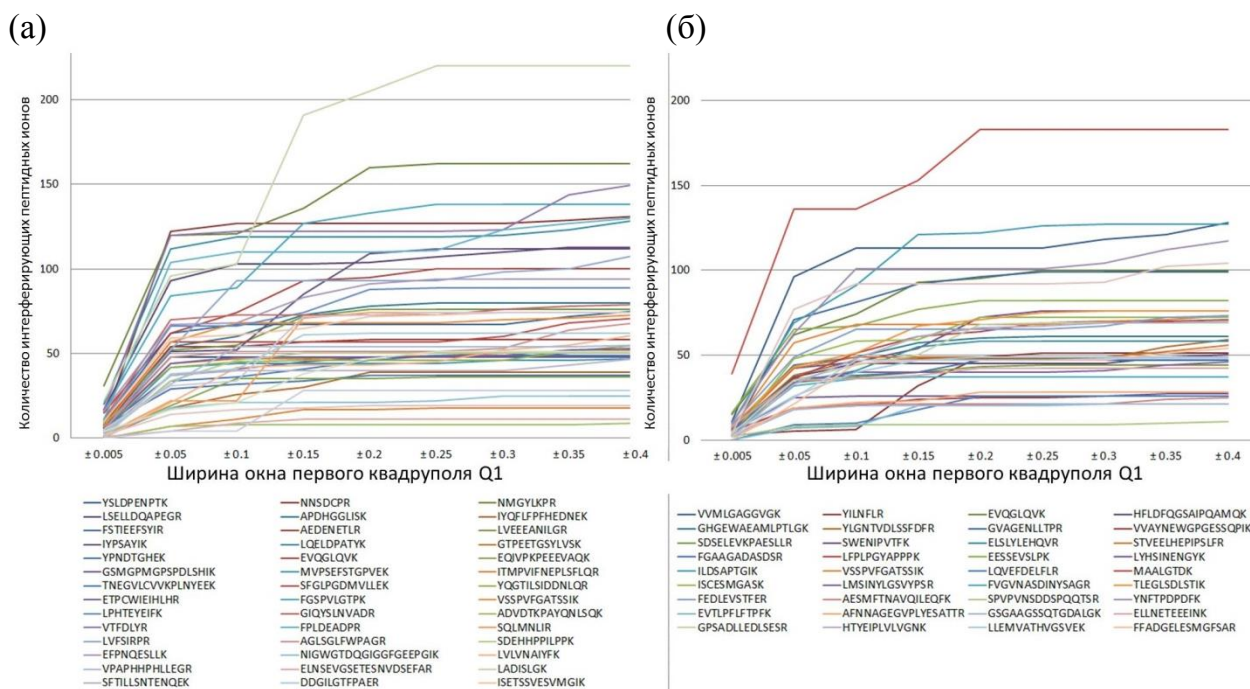


Рисунок 19. Зависимость количества интерферирующих пептидных ионов от ширины окна первого квадруполя Q1 для пептидов контрольной выборки (а) и случайного набора (б). График построен на основе данных, полученных с использованием ПО SRM Collider (см. раздел 3.9).

На рисунке 19а для пептидов, вошедших в состав результирующей выборки, показано, как меняется количество интерферирующих пептидных ионов в зависимости от ширины окна первого квадруполя Q1. Наибольшее количество потенциальных интерферирующих пептидных ионов (более 200) соответствует пептиду LADISLGK: чем меньше длина пептида, тем меньше вероятность того, что его фрагменты окажутся уникальными в рамках всего протеома человека.

При анализе рисунка 19 обращает внимание тот факт, что с точки зрения количества интерферирующих пептидных ионов контрольная и случайная группы пептидов практически не отличаются. Следовательно, интерференцию следует ожидать на уровне фрагментов, возникающих в результате диссоциативных переходов пептидных ионов. Количество потенциальных интерферирующих фрагментов зависит от ширины окна первого, так и второго квадруполя. На

рисунке 20 показана зависимость суммарного количества интерферирующих фрагментов для трех диссоциативных переходов, по которым был детектирован пептид GVMASSLQELISK белка O60543 (активатор клеточной смерти).

С использованием инструмента SRM Collider изучались два варианта интерференции. Интерференция между пептидными ионами показана на рис. 18, тогда как на рис. 20 рассматривается интерференция фрагментов, возникающая в результате диссоциативных переходов.

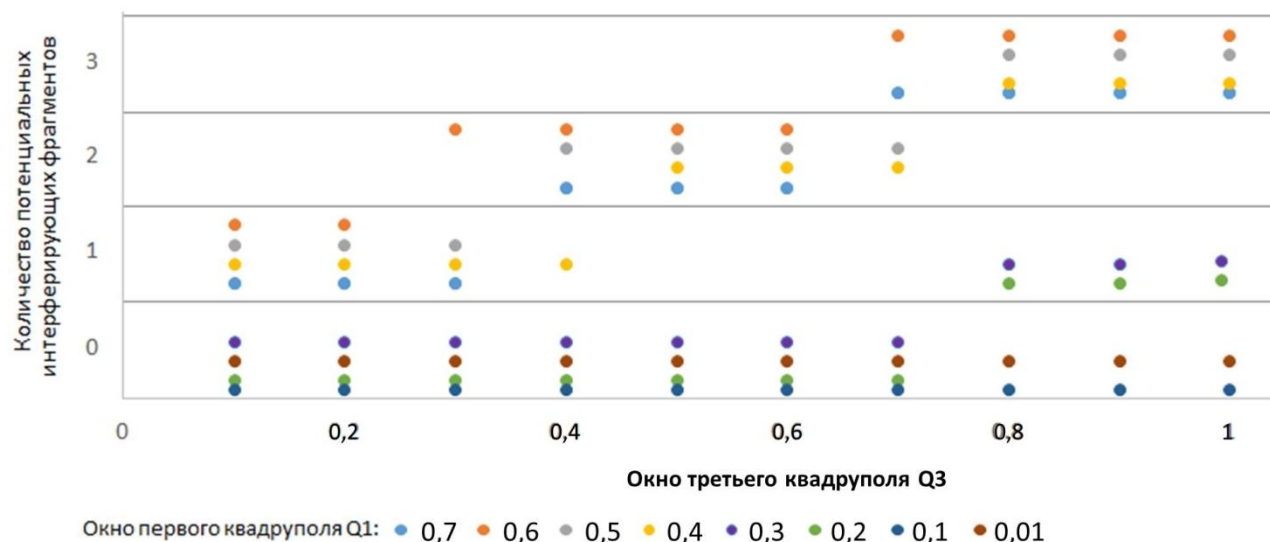


Рисунок 20. Интерференция Q3: зависимость суммарного количества потенциальных интерферирующих фрагментов для трех переходов пептида GVMASSLQELISK (белок O60543) от ширины окна первого квадруполя Q1 и третьего квадруполя Q3.

При высокой точности детекции масс-зарядных характеристик ионов (ширина окна Q1 0-0,3 и Q3 на уровне 0-0,7) вероятность возникновения интерференции мала (рис. 20). В то же время, подобные настройки существенно снижают производительность измерений и поэтому не применяются при хромосомотцентричном подходе.

Сравнительный анализ количества интерферирующих ионов для парных (т.е. относящихся к продукту одного гена) пептидов, вошедших в различные сегменты экспериментальной выборки, проводили для оценки влияния интерференции на разницу в результатах измерений концентраций парных пептидов. Вне зависимости от того, к какому сегменту экспериментальной выборки относился пептид,

количество потенциальных интерферирующих пептидных ионов в среднем составляло 30.

Исключением являлся пептид VTFDLYR, состоящий всего из 7 аминокислотных остатков. Для остальных пептидов количество интерферирующих фрагментов для 3 переходов, по которым был детектирован пептид не превышало 15, в среднем составляя $5 \pm$.

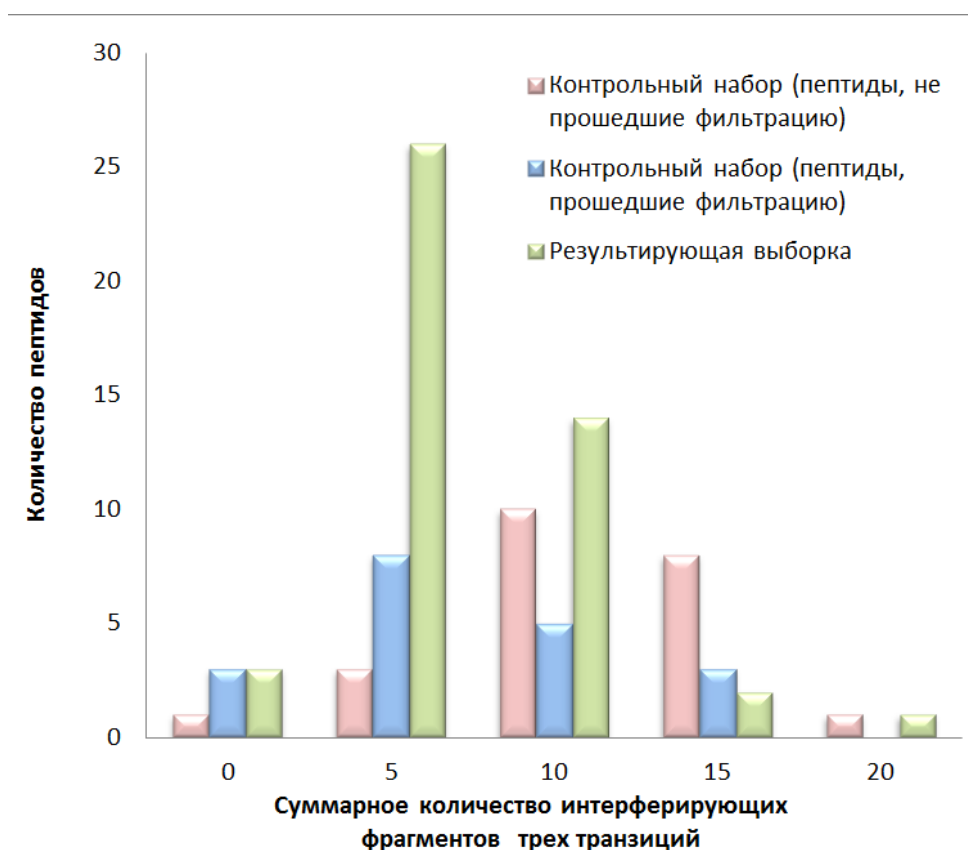


Рисунок 21. Количество потенциальных интерферирующих фрагментов для трех переходов: (а) пептидов контрольной выборки; (б) пептидов из случайного набора, удовлетворявших установленным фильтрам; (в) пептидов из случайного набора, не удовлетворявших установленным критериям.

Для пептидов, которые были отобраны на основе оценки воспроизводимости в технических повторах, количество потенциальных интерферирующих пептидных ионов и фрагментов в среднем превышает аналогичное у пептидов с высокой воспроизводимостью. Например, для пептида DLHDSFDFHGR коэффициент вариации интенсивностей фрагментов в технических повторах составил менее 12%. Для него существуют 40 интерферирующих пептидных ионов и два

интерферирующих фрагмента для трех фрагментов, использованных при детекции. Это количество интерферирующих фрагментов в среднем соответствует данным, полученным для пептидов из контрольной выборки. В то же время, для парного пептида VFGWVHR существуют 24 интерферирующих пептидных ионов и девяти интерферирующих фрагментов. Коэффициент вариации интенсивности фрагментов этого пептида в технических повторах колебался в пределах 35–68%. Исходя из описанного выше, можно заключить, что в формировании интерферирующего сигнала наибольшую роль играют фрагменты, детектируемые в третьем квадруполе.

Таким образом, предварительный отбор фрагментных ионов при планировании эксперимента существенно влияет на достоверность полученных результатов и, как следствие, на возможность их использования в обнаружении биологических маркеров.

Из анализа вышеприведенных графиков (рис.19, рис 20, рис. 21) видно, что SRM Collider позволяет анализировать свойства ионов, связанные исключительно с аминокислотным составом. Поэтому для получения биологически осмысленных закономерностей необходимо сопоставлять данные каскадной фильтрации с дополнительными данными, имеющими отношение к свойствам белков как биологических объектов.

4.4.3. Применение базы данных в научной работе

С целью продемонстрировать применение базы данных протеотипических пептидов в решении научных задач был проведен анализ результирующей выборки на предмет наличия белков, для которых известны сплайс-варианты (табл. 7).

Сравнение детектированных пептидных последовательностей, вошедших в состав результирующей выборки, с последовательностями сплайс-вариантов белков показало, что четыре протеотипических пептида относятся к сплайс-опосредованным вариантам разных белков. Выявлено, что если хотя бы один из двух детектированных пептидов одного белка картируется на сплайс-формы других белков, то разница между концентрациями таких пептидов составляет два и более порядков (см. таблицу 1).

Например, пептид LAVNMVPFPR бета-тубулина 8 A6NNZ2 (Tubulin beta-8 chain-like protein LOC260334) входит в состав сплайс-формы третьей субъединице бета-тубулина (Q40665-2, Tubulin beta-3 chain). Разница с концентрацией парного пептида HGCYLTVA AIFR составляет два порядка. Показательным является пептид LVLVNAIYFK серпина B4 (P48594, SERPINB4), соответствующий сплайс-формам трех белков человека: интерлейкин связывающего фактора (Q5RFJ1-2 Interleukin enhancer-binding factor 2), серпина B5 (P36952-2 Serpin B5) и серпина B8 P50452-3 (Serpin B8). Разница его концентрации с парным пептидом VLHFDQVTENTTEK составила более 6 порядков. При этом количество идентификаций панорамными масс-спектрометрическими методами пептида LVLVNAIYFK более чем в 5 раз превышает количество идентификаций пептида VLHFDQVTENTTEK.

Показательно, что по данным репозитория PRIDE пептид бета-тубулина LAVNMVPFPR, был идентифицирован панорамными масс-спектрометрическими методами 178 раз, как в крови, так и в нормальных и патологических тканях. При этом, пептид HGCYLTVA AIFR того же белка не был идентифицирован ранее.

Для сравнения провел анализ количества экспериментов, в которых были детектированы пептиды, картирующиеся на несколько сплайс-опосредованных вариантов белков. Было получено, что для пептидов LAVNMVPFPR и LVLVNAIYFK разница в количестве идентификации для парных пептидов составляла несколько раз. Кроме того, количество экспериментов, в которых были идентифицированы эти пептиды, превышало количество экспериментов, в которых был идентифицирован белок, на 89 и 20 экспериментов соответственно.

Следовательно, идентификация этих пептидов в биологическом материале может свидетельствовать о наличии в нем не одного, а нескольких белков.

Выявлены также примеры, когда пептиды одного белка картировались на разные сплайс-опосредованные формы одного белка. В этом случае концентрации парных пептидов были в пределах одного порядка. Например, пептиды GTPEETGSYLVS K и LSELLDQAPEGR транслоцируемого при лимфомах белка Q9UDY8 (Mucosa-associated lymphoid tissue lymphoma translocation protein 1)

соответствуют как его канонической (мастерной) форме, так и его сплайс-варианту Q9UDY8-2. Кроме того, пептид GTPEETGSYLVSK картировался еще на одну сплайс-форму того же белка (Q9UDY8-3). Таким образом, видно, что качество выполнения этапа планирования эксперимента – в том числе проверка уникальности исследуемых пептидных последовательностей отражается на конечном результате количественных измерений. Использование результатов измерений неуникальных пептидных последовательностей возможно в случае сочетания их с измерениями содержания пептидов, уникальных для белка или его сплайс-опосредованной формы.

Таблица 7. Соответствие протеотипических пептидов из результирующей выборки сплайс-опосредованным вариантам белков

Название (UniProt AC)	Аминокислотная последовательность пептида	Сплайс-опосредованные формы (Uniprot AC)	Δ^*
Бета-тубулин 8 (A6NNZ2)	HGCVLTVAAlFR		10^2
	LAVNMVPFPR	Третья субъединица бета-тубулина (Q40665-2)	
Оксистерол-связывающий белок (Q9BXW6)	LQELDPATYK	Анкирин домен-содержащий белок (Ankyrin repeat domain-containing protein Q6UB98-2)	10^3
	NDFSIWSILR	Оксистерол-связывающий белок (Q9BXW6-4), (Q91XL9-2)	
Серпин B4 (P48594)	LVLVNAIYFK	Интерлейкин связывающий фактор (Q5RFJ1-2), серпин B5 (P36952-2), серпин B8 (P50452-3)	10^6
	VLHFDQVTENTTEK		

* Δ - отношение между концентрациями парных пептидов, относящихся к одному белку.

В состав экспериментальной выборки вошли 14 белка, пептиды которых соответствовали исключительно канонической форме целевого белка, аминокислотные последовательности этих пептидов не совпадали с известными в масштабах протеома сплайс-опосредованными формами. Для таких белков

измеренные концентрации парных пептидов отличались не более, чем на 100%. Таким образом, пример с выявлением сплайс-опосредованных вариантов белков показывают возможность использования хромосомотцентричной базы данных протеотипических пептидов в контексте решения научной задачи. В основе формирования базы данных лежит метод, показанный на рисунке 22. Разработанная информационная структура позволяет формировать систему запросов к экспериментальным данным. На основе запросов конструируются каскадные фильтры, и проводится сегментирование массива экспериментов. Применение фильтров к сегментам информационного массива обеспечивает формирование фокусированной выборки белков. Выборка, будучи ограниченной по масштабу, в то же время наследует информационные свойства исходной хромосомы. Анализ выборки белков позволяет планировать дальнейшие исследования, например, основанные на направленном масс-спектрометрическом измерении сплайс-опосредованных форм белков [Mirza и др., 2008].

На рисунке 22 представлена блок-схема процесса обработки результатов детекции протеотипических пептидов, кодируемых хромосомой 18 белков. В качестве инструмента оценки воспроизводимости и достоверности результатов детекции белков в биологическом материале использовалась программная оболочка для ввода информации о протеотипических пептидах, а также пакет подпрограмм, предназначенных для аналитической обработки вводимых данных. Интерпретация результатов масс-спектрометрических экспериментов, выполненных методом мониторинга диссоциативных переходов, позволила оценить производительность исследований. Результатом протеомного исследования белков хромосомы 18 в биологическом материале стала выборка белков, детектированных с высокой вероятностью, а также выборка белков, количественные измерения которых требуют проверки и уточнения в дальнейших экспериментах.

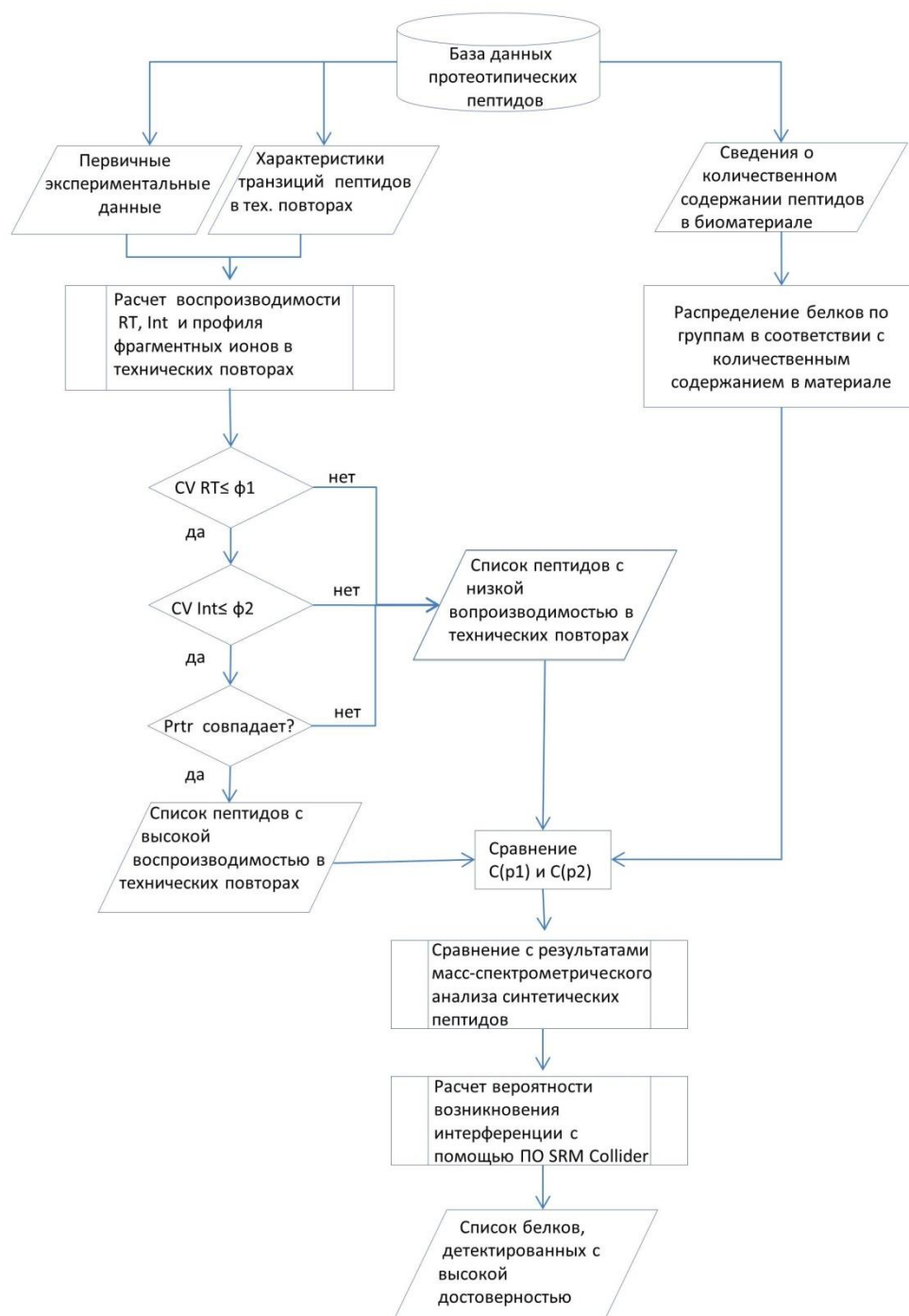


Рисунок 22. Блок-схема для оценки воспроизводимости и достоверности результатов детекции белков в биологическом материале методом мониторинга диссоциативных переходов. Обозначения: RT – время удержания, усредненное для всех фрагментных ионов в техническом повторе; Int – интенсивность каждого фрагментного иона в техническом повторе; Prtr – последовательность фрагментных ионов, отсортированных по значению интенсивности в порядке убывания; C(p1) – результат количественного измерения пептида №1 белка AC, C(p2) – результат количественного измерения пептида №2 белка AC, где AC – идентификатор белка в базе данных UniProt.

5. ЗАКЛЮЧЕНИЕ

Современные технологии позволяют получать экспериментальные данные в высокопроизводительном режиме. Интерпретация этих данных, извлечение из них биологического смысла затруднено в связи с отсутствием «золотого стандарта». Отсутствует статистическая модель, позволяющая проводить оценку качества экспериментальных данных в автоматическом режиме вне зависимости от использованных приборов, протоколов пробоподготовки и ПО, использованного для обработки [Orchard, Hermjakob, 2008].

С точки зрения традиционной молекулярно-биологической науки сама постановка задачи инвентаризации протеома человека в хромосомотцентричном формате является неестественной. Неясно, каким образом инвентаризация белков одной хромосомы может способствовать решению фундаментальных проблем, связанных с описанием молекулярных механизмов действия клетки [Archakov и др., 2012]. Предположение, что биомаркеры, связанные с тем или иным заболеванием, в большей степени тяготеют к локализации соответствующих генов на одной хромосоме, до сих пор не нашло убедительных подтверждений. В данной работе хромосомотцентричный подход к протеомным исследованиям рассматривался не как научно-исследовательская, а как технологическая задача, без решения которой дальнейшее развитие белковой химии не представляется возможным.

Проектирование исследования требует определения свойств обрабатываемого «сырья» и разработки информационных и функциональных структур, описывающих этапы, информационные потоки и контрольные точки. В данной работе была разработана информационная структура данных протеомного исследования, посредством которого возможно осуществление контроля качества результатов выполнения технологических операций.

Протеомное исследование, выполняемое методом направленной масс-спектрометрии, включает биохимические, физические, методические и информационные этапы [Orchard, Hermjakob, 2008]. Исходным «сырьем» является физический объект «биологический материал», тогда как результатом является информационный объект, описывающий характеристики исходного объекта.

Особенностью исследования, основанного на масс-спектрометрическом методе МДП является отсутствие промежуточных результатов, что не позволяет контролировать все стадии. Подготовительные этапы исследования, такие как биоинформатическое планирование эксперимента, пробоподготовка, синтез не производят информационного результата, а только трансформируют исходный объект. Поэтому оценка результатов протеомного исследования может производиться только постфактум. С точки зрения моделирования исследовательского процесса выявление этапов, изменение в протоколах которых приводит к созданию неудовлетворительного результата, является крайне важным, так как целью любого исследования является увеличение количества правдоподобных, корректных результатов по отношению к ошибочным. Первостепенной задачей является построение системы оценки полученных результатов, что, в случае направленной масс-спектрометрии в области протеомики, на сегодняшний день является нерешенной задачей.

В данной работе была проведена формализации результатов масс-спектрометрических измерений, проведенных на базе массива экспериментальных данных, накопленных в ходе выполнения российской части проекта «Протеом человека» [Ропомarenko и др., 2014; Zgoda и др., 2013; Пономаренко и др., 2015].

Система запросов к разработанной информационной структуре данных протеомного исследования методом мониторинга диссоциативных переходов позволила сегментировать полученные результаты на группы с высокой, средней и низкой степенью воспроизводимости. Поскольку протеомное исследование представляет собой не меняющийся во времени набор операций, то из общих соображений следует, что проведение одних и тех же операций над одним и тем же объектом должно проводить к одному и тому же результату. Получаемые на практике результаты различаются в 90% случаев, что свидетельствует о наличии в исследовательском процессе технологических операций, которые протоколированы с недостаточной степенью детализации. С точки зрения решения фундаментальной научной задачи поиска биологических маркеров, оценка достоверности результатов измерений является первостепенной задачей.

В ряде случаев показано, что существующие биоинформатические инструменты облегчают задачу выявления воспроизводимых результатов из противоречивого массива данных, накопленных мировым сообществом [Farrah и др., 2012; Reiter и др., 2011; Surinova и др., 2013]. Однако область их применения небольшая и зачастую устанавливает жесткие требования к типу используемых данных и экспериментальным протоколам. Поэтому массив накопленных на сегодняшний день сведений о детекции белков методом мониторинга диссоциативных переходов не отражает реальных молекулярно-биологических процессов, а является сводом субъективных оценок исследователей. В текущем виде использование этих данных с целью установления молекулярно-биологических закономерностей невозможно. В результате оценки воспроизводимости результатов детектирования белков, кодируемых генами хромосомы 18 методом мониторинга диссоциативных переходов, выявлено в общей сложности 46 пептидов, соответствующих 23 белкам, кодируемым генами хромосомы 18 человека, пригодных для проведения количественных измерений в биологическом материале с высокой степенью воспроизводимости результатов. Воспроизводимость результатов в технических повторах позволяет оценить возможность использования данных, полученных различными исследователями и выявлять биологически важные результаты. Интерпретация экспериментальных данных позволяет перейти от эпизодических наблюдений белков в тканях к формированию стабильной аннотации биологического материала с точки зрения его протеомного профиля. Воспроизводимость детекции белка позволяет судить о стабильности его существования на протеомном уровне, а также о возможности его использования в качестве биологического маркера.

Наибольший интерес с точки зрения установления биологически значимых закономерностей и поиска клинически значимых маркеров представляет возможность выявления форм белков методами направленной протеомики [Lisitsa и др., 2014; Wu и др., 2012].

Для иллюстрации возможностей практического применения предлагаемого подхода было проведено исследование сплайс-опосредованных вариантов белков

человека. На массиве данных, являющимся протеомом отдельно взятой хромосомы, были выявлены четыре пептида, входящих в состав сплайс-опосредованных вариантов двух и более белков человека.

Предложенная в данной работе методология, рассматривающая протеомное исследование, позволяет предположить наличие взаимосвязи между критерием воспроизводимости результатов и интерференцией пептидных ионов - одной из возможных причин, по которой масс-спектрометрии не удастся выявить биомаркеры [Blonder, Veenstra, 2007].

6. ВЫВОДЫ

1. Разработана структура данных, описывающая исследования протеома человека методом направленной количественной масс-спектрометрии (методом мониторинга диссоциативных переходов). На основе разработанной структуры реализована база данных протеотипических пептидов белков, кодируемых генами хромосомы 18.
2. Показано, что среди свойств протеотипических пептидов критическими параметрами, отвечающими за воспроизводимость результатов при количественных измерениях являются совпадение профиля транзиций в технических повторах, расхождение интенсивности фрагментных ионов в которых не превышает 20%.
3. В результате оценки результатов протеомного профилирования белков, определено в общей сложности 23 белка и соответствующие им 46 протеотипических пептидов, которые могут быть использованы в качестве стандартов для проведения количественных измерений этих белков в различных типах биологического материала.
4. Аннотация полученной выборки протеотипических пептидов с использованием биоинформатических ресурсов показала, что точность количественного измерения белков зависит от специфичности пептида для конкретной сплайс-опосредованной формы белка. Выявлены пептиды, подтверждающие наличие в клеточной линии HepG2 сплайс-опосредованных форм анкирин домен-содержащий белка (Q6UB98-2) (LAVNMVPFPR), третьей субъединицы бета-тубулина (Q40665-2) (LQELDPATYK), интерлейкин связывающего фактора (Q5RFJ1-2), серпина B5 (P36952-2), серпина B8 (P50452-3) (LVLVNAIYFK).
5. Для белка FAM44C, участвующего в процессе формирования веретена деления клеток, впервые было подтверждено существование на протеомном уровне в клеточной линии HepG2.

7. СПИСОК СОКРАЩЕНИЙ

а.о. — аминокислотный остаток;

БД – база данных;

ВЭЖХ – высокоэффективная жидкостная хроматография;

ПО – программное обеспечение;

СО-вариант белка — сплайс-опосредованный вариант белка;

АС – UniProtKB;

FDA – Food & Drug Administration, управление по санитарному надзору за качеством пищевых продуктов и медикаментов, США;

GO – Gene Ontology;

HUPO – международная организация Протеом человека;

LC-MS/MS — жидкостная tandemная хромото-масс-спектрометрия;

MS/MS — tandemная масс-спектрометрия;

PPD – Plasma Proteome Database;

QQQ – масс-спектрометр с тройным квадруполом;

SRM – selected reaction monitoring, метод мониторинга диссоциативных переходов.

8. ОСНОВНЫЕ ПОНЯТИЯ И ТЕРМИНЫ

Время удержания: время пребывания исследуемого вещества в хроматографе.

Практически время удерживания определяют от момента ввода пробы вещества в хроматограф до момента регистрации максимума соответствующего хроматографического пика.

Группа пиков хроматограммы: группа коэлюирующих пиков

Интенсивность (высота) пика, h: расстояние от максимума пика до его основания, измеренное вдоль оси отклика детектора.

Коэлюция: одновременный выход веществ с хроматографической колонки.

Масс-спектр: то зависимость интенсивности ионного тока (количества вещества) от отношения массы к заряду (природы вещества).

Метод мониторинга диссоциативных переходов: метод направленной масс-спектрометрии, который заключается в детекции пептидного иона заданной массы и заряда, а также детекции нескольких его фрагментов заданной массы и заряда.

Направленная масс-спектрометрия (информационно-независимый режим) (в англоязычной литературе targeted): масс-спектрометрический подход, при котором регистрируется сигнал только от заранее заданных исследователем соединений.

Панорамная масс-спектрометрия (информационно-зависимый режим): масс-спектрометрический подход, при котором регистрируется сигнал от всех компонентов биологического образца.

Пептидный ион (родительский ион, прекурсорный ион): ионизированный пептид.

Переход (транзикация): пара пептидного иона и иона фрагмента.

Пик хроматограммы (хроматографический пик): участок хроматограммы, соответствующий площади ограниченной функцией хроматограммы в момент выхода определяемого вещества из колонки и базовой линией. Нулевая (базовая) линия хроматограммы - линия, соответствующая нулевой концентрации анализируемых веществ в элюате.

Площадь пика: площадь хроматограммы, заключенная между пиком и его основанием.

Протеотипический пептид: пептид с уникальной в рамках генома одного организма последовательностью, обладающий способностью, быть детектированным в масс-спектрометре и концентрация которого коррелирует с концентрацией белка.

Профиль фрагментных ионов: порядок фрагментов пептида, отсортированный по убыванию их интенсивностей.

Сигнал-шум: отношение уровня сигнала к уровню шума.

Технический повтор: однократное измерение содержания белка в образце.

Триптический пептид: пептид, полученный путем гидролиза белковой смеси трипсином.

Хроматограмма: кривая, изображающая зависимость концентрации соединений, выходящих из колонки с потоком подвижной фазы, от времени с момента начала разделения.

Целевой пептид: протеотипический пептид, отобранный для анализа методом мониторинга диссоциативных переходов.

Эксперимент: совокупность нескольких технических повторов, проведенных в одинаковых условиях на одном и том же образце.

m/z (масс-зарядная характеристика): отношение массы к заряду иона.

9. СПИСОК ЛИТЕРАТУРЫ

1. Abbatiello S.E. и др. Design, Implementation and Multisite Evaluation of a System Suitability Protocol for the Quantitative Assessment of Instrument Performance in Liquid Chromatography-Multiple // *Mol. Cell. Proteomics*. 2013. С. 2623–2639.
2. Addona T. а и др. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. // *Nat. Biotechnol.* 2009. Т. 27. № 7. С. 633–41.
3. Addona T.A. и др. A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease // *Nat. Biotechnol.* 2012. Т. 29. № 7. С. 635–643.
4. Aebersold R., Mann M. Mass spectrometry-based proteomics. // *Nature*. 2003. Т. 422. № 6928. С. 198–207.
5. Altschul S.F. и др. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs // *Nucleic Acids Res.* 1997. Т. 25. № 17. С. 3389–3402.
6. Alves G., Ogurtsov A.Y., Yu Y.-K. Assigning statistical significance to proteotypic peptides via database searches // *J Proteomics*. 2011. Т. 74. № 2. С. 199–211.
7. Anderson N.L. и др. The human plasma proteome: a nonredundant list developed by combination of four separate sources. // *Mol. Cell. Proteomics*. 2004. Т. 3. № 4. С. 311–326.
8. Anderson N.L. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. // *Clin. Chem.* 2010. Т. 56. № 2. С. 177–85.
9. Anderson N.L., Anderson N.G. The human plasma proteome: history, character, and diagnostic prospects. // *Mol. Cell. Proteomics*. 2002. Т. 1. № 11. С. 845–867.
10. Archakov A. и др. Gene-centric view on the human proteome project: The example of the Russian roadmap for chromosome 18 // *Proteomics*. 2011. Т. 11. № 10. С. 1853–1856.
11. Archakov A. и др. Chromosome-centric approach to overcoming bottlenecks in the Human Proteome Project // *Expert Rev Proteomics*. 2012. Т. 9. № 6. С. 667–676.
12. Bairoch a. The ENZYME database in 2000. // *Nucleic Acids Res.* 2000. Т. 28. № 1. С. 304–305.
13. Bao Y. и др. Detection and correction of interference in SRM analysis. // *Methods*. 2013. Т. 61. № 3. С. 299–303.

14. Berna M. и др. Strategic use of immunoprecipitation and LC/MS/MS for trace-level protein quantification: Myosin light chain 1, a biomarker of cardiac necrosis // *Anal. Chem.* 2007. Т. 79. № 11. С. 4199–4205.
15. Berna M. и др. Quantification of NTproBNP in rat serum using immunoprecipitation and LC/MS/MS: A biomarker of drug-induced cardiac hypertrophy // *Anal. Chem.* 2008. Т. 80. № 3. С. 561–566.
16. Blonder J., Veenstra T.D. Computational prediction of proteotypic peptides. // *Expert Rev. Proteomics.* 2007. Т. 4. № 3. С. 351–4.
17. Bondarenko P. V., Chelius D., Shaler T. a. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography - Tandem mass spectrometry // *Anal. Chem.* 2002. Т. 74. № 18. С. 4741–4749.
18. Camacho C. и др. BLAST+: architecture and applications. // *BMC Bioinformatics.* 2009. Т. 10. С. 421.
19. Carr S. а и др. Targeted Peptide Measurements in Biology and Medicine: Best Practices for Mass Spectrometry-based Assay Development Using a Fit-for-Purpose Approach. // *Mol. Cell. Proteomics.* 2014. С. 907–917.
20. Cham J. а и др. MRMAid-DB: a repository of published SRM transitions. // *J. Proteome Res.* 2010. Т. 9. № 1. С. 620–5.
21. Chelius D., Bondarenko P. V. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. // *J. Proteome Res.* 2002. Т. 1. № 4. С. 317–323.
22. Chen C. и др. Screening of missing proteins in the human liver proteome by improved MRM-approach-based targeted proteomics // *J. Proteome Res.* 2014. Т. 13. № 4. С. 1969–1978.
23. Craig R., Cortens J.P., Beavis R.C. The use of proteotypic peptide libraries for protein identification // *Rapid Commun. Mass Spectrom.* 2005. Т. 19. № 13. С. 1844–1850.
24. Crasto C. и др. MRMPATH and MRMutation, facilitating discovery of mass transitions for proteotypic peptides in biological pathways using a bioinformatics approach // *Adv. Bioinformatics.* 2013. Т. 2013.
25. Cummings J.L. Clinical evaluation as a biomarker for Alzheimer’s disease. // *J. Alzheimers. Dis.* 2005. Т. 8. С. 327–337.

26. Desiere F. и др. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. // *Genome Biol.* 2005. Т. 6. № 1. С. R9.
27. Desiere F. и др. The PeptideAtlas project. // *Nucleic Acids Res.* 2006. Т. 34. № Database issue. С. D655–D658.
28. Deutsch E. mzML: A single, unifying data format for mass spectrometer output // *Proteomics.* 2008. Т. 8. № 14. С. 2776–2777.
29. Deutsch E. и др. TraML--A Standard Format for Exchange of Selected Reaction Monitoring Transition Lists // *Mol. Cell. Proteomics.* 2012. Т. 11. № 4. С. R111.015040–R111.015040.
30. Domon B. Considerations on selected reaction monitoring experiments: Implications for the selectivity and accuracy of measurements. // *Proteomics. Clin. Appl.* 2012. Т. 6. № 11-12. С. 609–14.
31. Duncan M.W., Aebersold R., Caprioli R.M. The pros and cons of peptide-centric proteomics. // *Nat. Biotechnol.* 2010. Т. 28. № 7. С. 659–664.
32. Fan J. и др. MRMaid 2.0: Mining PRIDE for Evidence-Based SRM Transitions // *Omi. A J. Integr. Biol.* 2012. Т. 16. № 9. С. 483–488.
33. Farrah T. и др. PASSEL: the PeptideAtlas SRMexperiment library. // *Proteomics.* 2012. Т. 12. № 8. С. 1170–5.
34. Farrah T. и др. The state of the human proteome in 2012 as viewed through PeptideAtlas // *J. Proteome Res.* 2013. Т. 12. № 1. С. 162–171.
35. Fortin T. и др. Multiple reaction monitoring cubed for protein quantification at the low nanogram/milliliter level in nondepleted human serum // *Anal. Chem.* 2009. Т. 81. № 22. С. 9343–9352.
36. Gallien S. и др. Highly multiplexed targeted proteomics using precise control of peptide retention time. // *Proteomics.* 2012. Т. 12. № 8. С. 1122–33.
37. Garcia-Blanco M.A., Baraniak A.P., Lasda E.L. Alternative splicing in disease and therapy. // *Nat. Biotechnol.* 2004. Т. 22. № 5. С. 535–546.
38. Gaudet P. и др. The neXtProt knowledgebase on human proteins: current status // *Nucleic Acids Res.* 2015. Т. 43. С. 764–770.
39. Geiger T. и др. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins // *Mol. Cell. Proteomics.* 2012. Т. 11.

40. Godoy L.M.F. de и др. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. // *Nature*. 2008. Т. 455. № 7217. С. 1251–1254.
41. Gorshkov a. V. и др. Liquid chromatography at critical conditions: comprehensive approach to sequence-dependent retention time prediction // *Anal. Chem.* 2006. Т. 78. № 22. С. 7770–7777.
42. Govorun V.M., Archakov a. I. Proteomic technologies in modern biomedical science // *Biochem.* 2002. Т. 67. № 10. С. 1109–1123.
43. Hall S.C. и др. Large-scale inter-laboratory study to develop, analytically validate and apply highly multiplexed, quantitative peptide assays to measure cancer-relevant proteins in plasma // *MCP*. 2015. Т. 14. № 9. С. 2357–2374.
44. Hüttenhain R. и др. Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. // *Sci. Transl. Med.* 2012. Т. 4. № 142. С. 142–194.
45. Jaquinod M. и др. Mass spectrometry-based absolute protein quantification: PSAQTM strategy makes use of «noncanonical» proteotypic peptides. // *Proteomics*. 2012. Т. 12. № 8. С. 1217–1221.
46. Jones P. и др. PRIDE: a public repository of protein and peptide identifications for the proteomics community. // *Nucleic Acids Res.* 2006. Т. 34. № Database issue. С. 659–663.
47. Jones P. и др. PRIDE: New developments and new datasets // *Nucleic Acids Res.* 2008. Т. 36. № SUPPL. 1. С. 878–883.
48. Keil B. Specificity of Proteolysis // Springer-Verlag Berlin, Ger. 1992.
49. Kelleher N.L. A cell-based approach to the human proteome project. // *J. Am. Soc. Mass Spectrom.* 2012. Т. 23. № 10. С. 1617–24.
50. Keshishian H. и др. Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. // *Mol. Cell. Proteomics*. 2009. Т. 8. № 10. С. 2339–2349.
51. Kiefer F. и др. The SWISS-MODEL Repository and associated resources // *Nucleic Acids Res.* 2009. Т. 37. № SUPPL. 1. С. 387–392.
52. Kim H. и др. Development of Biomarkers for Screening Hepatocellular Carcinoma Using Global Data Mining and Multiple Reaction Monitoring // *PLoS One*. 2013. Т. 8. № 5.

53. Kim M.-S. и др. A draft map of the human proteome // Nature. 2014. Т. 509. № 7502. С. 575–581.
54. Kim Y.J. и др. Verification of the Biomarker Candidates for Non-small-cell Lung Cancer Using a Targeted Proteomics Approach // J. Proteome Res. 2015. С. 1412–1419.
55. Koenig T. и др. Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics // J. Proteome Res. 2008. Т. 7. № 9. С. 3708–3717.
56. Kopylov A.T. и др. Combined use of irreversible binding and MRM technology for low- and ultralow copy-number protein detection and quantitation // Proteomics. 2013. Т. 13. № 5. С. 727–742.
57. Krokhin O. V., Spicer V. Predicting peptide retention times for proteomics // Curr. Protoc. Bioinforma. 2010. № SUPPL. 31. С. 1–15.
58. Kusebauch U. и др. Using PeptideAtlas, SRMAtlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics. // Curr. Protoc. Bioinformatics. 2014. Т. 46. № June.
59. Kuster B. и др. Scoring proteomes with proteotypic peptide probes. // Nat. Rev. Mol. Cell Biol. 2005. Т. 6. № 7. С. 577–583.
60. Kuzyk M. а и др. Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. // Mol. Cell. Proteomics. 2009. Т. 8. № 8. С. 1860–77.
61. Lane L. и др. Metrics for the Human Proteome Project---2013-2014 and Strategies for Finding Missing Proteins // 2014. Т. 13. № 1. С. 15–20.
62. Lange V. и др. Selected reaction monitoring for quantitative proteomics: a tutorial. // Mol. Syst. Biol. 2008a. Т. 4. № 222. С. 222.
63. Lange V. и др. Selected reaction monitoring for quantitative proteomics: a tutorial. // Mol. Syst. Biol. 2008b. Т. 4. № 222. С. 1–14.
64. Lange V. и др. Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. // Mol. Cell. Proteomics. 2008c. Т. 7. С. 1489–1500.
65. Lisitsa A. и др. Profiling proteoforms: promising follow-up of proteomics for biomarker discovery. // Expert Rev. Proteomics. 2014. Т. 11. № 1. С. 121–129.
66. Liu H. и др. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. // Anal. Chem. 2004. Т. 76. № 14. С. 4193–201.

67. Liu H., Sadygov R.G., Yates J.R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. // *Anal. Chem.* 2004. T. 76. № 14. С. 4193–201.
68. Liu T. и др. Analysis of Serum Total and Free PSA Using Immunoaffinity Depletion Coupled to SRM: Correlation with Clinical Immunoassay Tests // *J Proteomics.* 2012. T. 75. № 15. С. 4747–4757.
69. Ludwig C. и др. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. // *Mol. Cell. Proteomics.* 2012. T. 11. № 3. С. M111.013987.
70. MacLean B. и др. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. // *Bioinformatics.* 2010. T. 26. № 7. С. 966–8.
71. Mallick P. и др. Computational prediction of proteotypic peptides for quantitative proteomics. // *Nat. Biotechnol.* 2007. T. 25. № 1. С. 125–31.
72. Martens L. и др. PRIDE: The proteomics identifications database // *Proteomics.* 2005. T. 5. № 13. С. 3537–3545.
73. Martin D.B. и др. MRMer, an interactive open source and cross-platform system for data extraction and visualization of multiple reaction monitoring experiments. // *Mol. Cell. Proteomics.* 2008. T. 7. № 11. С. 2270–8.
74. Mirza M. и др. Osteopontin-c is a selective marker of breast cancer. // *Int. J. Cancer.* 2008. T. 122. № 4. С. 889–97.
75. Mohammed Y. и др. PeptidePicker: A scientific workflow with web interface for selecting appropriate peptides for targeted proteomics experiments // *J. Proteomics.* 2014. T. 106. С. 151–161.
76. Mörtstedt H. и др. Targeted Proteomic Analyses of Nasal Lavage Fluid in Persulfate-Challenged Hairdressers with Bleaching Powder-Associated Rhinitis // *J. Proteome Res.* 2015. T. 14. № 2. С. 860–873.
77. Naegle K., Gough N.R., Yaffe M.B. Criteria for biological reproducibility : What does “ n ” mean ? // *Sci. Signal.* 2015. T. 8. № 371. С. 8–10.
78. Nanjappa V. и др. Plasma Proteome Database as a resource for proteomics research : 2014 update // *Nucleic Acids Res.* 2014. T. 42. № December 2013. С. 959–965.
79. Nedelkov D. Population proteomics: Investigation of protein diversity in human populations // *Proteomics.* 2008. T. 8. С. 779–786.

80. Neilson K. a. и др. Less label, more free: Approaches in label-free quantitative mass spectrometry // *Proteomics*. 2011. Т. 11. № 4. С. 535–553.
81. Nesvizhskii A.I., Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. // *Mol. Cell. Proteomics*. 2005. Т. 4. № 10. С. 1419–1440.
82. Olsen J. V, Ong S.-E., Mann M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. // *Mol. Cell. Proteomics*. 2004. Т. 3. № 6. С. 608–614.
83. Omenn G.S. The strategy, organization, and progress of the HUPO Human Proteome Project // *J. Proteomics*. 2014. Т. 100. С. 3–7.
84. Orchard S., Hermjakob H. The HUPO proteomics standards initiative - Easing communication and minimizing data loss in a changing world // *Brief. Bioinform.* 2008. Т. 9. № 2. С. 166–173.
85. Orchard S., Ping P. HUPO World Congress Publication Committee meeting. August 2008, Amsterdam, The Netherlands. // *Proteomics*. 2009. Т. 9. № 3. С. 502–503.
86. Paik Y.-K. и др. Standard guidelines for the chromosome-centric human proteome project. // *J. Proteome Res.* 2012. Т. 11. № 4. С. 2005–13.
87. Patterson S.D., Aebersold R.H. Proteomics: the first decade and beyond. // *Nat. Genet.* 2003. Т. 33. № March. С. 311–323.
88. Pearson W.R., Lipman D.J. Improved tools for biological sequence comparison. // *Proc. Natl. Acad. Sci. U. S. A.* 1988. Т. 85. № 8. С. 2444–2448.
89. Pearsons L. and S.D. The Human Genome Project: a paradigm for information management in the life sciences // *Faseb J.* 1991. Т. 5. С. 35–39.
90. Percy A.J. и др. Method and platform standardization in MRM-based quantitative plasma proteomics // *J. Proteomics*. 2013a. Т. 95. С. 66–76.
91. Percy A.J. и др. Method and platform standardization in MRM-based quantitative plasma proteomics // *J. Proteomics*. 2013b. Т. 95. С. 66–76.
92. Perkins D.N. и др. Probability-based protein identification by searching sequence databases using mass spectrometry data // *Electrophoresis*. 1999. Т. 20. № 18. С. 3551–3567.
93. Picotti P. и др. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes // *Nat. Methods*. 2010. Т. 7. № 1. С. 43–48.
94. Picotti P., Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. // *Nat. Methods*. 2012. Т. 9. № 6. С. 555–66.

95. Pirmoradian M. и др. Rapid and deep human proteome analysis by single-dimension shotgun proteomics. // *Mol. Cell. Proteomics*. 2013. Т. 12. № 11. С. 3330–8.
96. Ponomarenko E. и др. Comparative ranking of human chromosomes based on post-genomic data. // *OMICS*. 2012. Т. 16. № 11. С. 604–11.
97. Ponomarenko E.A. и др. Chromosome 18 transcriptoproteome of liver tissue and HepG2 Cells and targeted proteome mapping in depleted plasma: Update 2013 // *J. Proteome Res*. 2014. Т. 13. № 1. С. 183–190.
98. Rappsilber J. и др. Large-Scale Proteomic Analysis of the Human Spliceosome // *Genome Res*. 2002. Т. 12. № 8. С. 1231–1245.
99. Reiter L. и др. mProphet: automated data processing and statistical validation for large-scale SRM experiments. // *Nat. Methods*. 2011. Т. 8. № 5. С. 430–5.
100. Rodriguez J. и др. Does Trypsin Cut Before Proline? // *J. Proteome Res*. 2007. Т. 7. № 1. С. 300–305.
101. Röst H., Malmström L., Aebersold R. A computational tool to detect and avoid redundancy in selected reaction monitoring // *MCP*. 2012. С. 1–23.
102. Sajic T., Liu Y., Aebersold R. Using Data-Independent , High Resolution Mass Spectrometry in Protein Biomarker Research : Perspectives and Clinical Applications Abbreviation list : // 2014. С. 1–29.
103. Sarkozy A. и др. Germline BRAF mutations in noonan, LEOPARD, and cardiofaciocutaneous Syndromes: Molecular diversity and associated phenotypic spectrum // *Hum. Mutat*. 2009. Т. 30. № 4. С. 695–702.
104. Schaab C. и др. Analysis of High Accuracy, Quantitative Proteomics Data in the MaxQB Database. // *Mol. Cell. Proteomics*. 2012. Т. 11. № 3. С. M111.014068.
105. Schmidt A., Picotti P., Aebersold R. Proteomanalyse und Systembiologie // *BioSpektrum*. 2008. Т. 14. С. 44–46.
106. Selevsek N. и др. Systematic quantification of peptides/proteins in urine using selected reaction monitoring. // *Proteomics*. 2011. Т. 11. № 6. С. 1135–47.
107. Shargunov A. V. и др. Tissue-specific alternative splicing analysis reveals the diversity of chromosome 18 transcriptome // *J. Proteome Res*. 2014. Т. 13. № 1. С. 173–182.
108. Sharma V. и др. Panorama: A Targeted Proteomics Knowledge Base. // *J. Proteome Res*. 2014.

109. Sherman J. и др. Unique ion signature mass spectrometry, a deterministic method to assign peptide identity. // *Mol. Cell. Proteomics*. 2009. Т. 8. № 9. С. 2051–62.
110. Sherman J., Molloy M.P., Burlingame A.L. Why complexity and entropy matter: information, posttranslational modifications, and assay fidelity. // *Proteomics*. 2012. Т. 12. № 8. С. 1147–50.
111. Sherrod S.D. и др. Label-free quantitation of protein modifications by pseudo selected reaction monitoring with internal reference peptides // *J. Proteome Res.* 2012. Т. 11. № 6. С. 3467–3479.
112. Sherwood C.A. и др. Transition List Assembly // 2010. Т. 8. № 10. С. 4396–4405.
113. Shi T. и др. Advancing the sensitivity of selected reaction monitoring-based targeted quantitative proteomics. // *Proteomics*. 2012. Т. 12. № 8. С. 1074–92.
114. Sigrist C.J. а и др. PROSITE, a protein domain database for functional characterization and annotation // *Nucleic Acids Res.* 2009. Т. 38. № SUPPL.1. С. 161–166.
115. Smith L.M., Kelleher N.L., Proteomics T.C. for T.D. Proteoform: a single term describing protein complexity // *Nat Methods*. 2013. Т. 10. № 3. С. 186–187.
116. Srebrow A., Kornblihtt A.R. The connection between splicing and cancer. // *J. Cell Sci.* 2006. Т. 119. С. 2635–2641.
117. Surinova S. и др. Automated selected reaction monitoring data analysis workflow for large-scale targeted proteomic studies. // *Nat. Protoc.* 2013. Т. 8. № 8. С. 1602–19.
118. Szklarczyk D. и др. The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored // *Nucleic Acids Res.* 2011. Т. 39. № SUPPL. 1. С. 561–568.
119. Tang H.-Y. и др. Rapid Verification of Candidate Serological Biomarkers Using Gel-based, Label-free Multiple Reaction Monitoring Hsin-Yao // *J Proteome Res.* 2011. Т. 10. № 9. С. 4005–4017.
120. Teleman J. и др. Automated selected reaction monitoring software for accurate label-free protein quantification. // *J. Proteome Res.* 2012. Т. 11. № 7. С. 3766–73.
121. Venter J.C. и др. The sequence of the human genome. // *Science*. 2001. Т. 291. № February. С. 1304–1351.
122. Whiteaker J.R. и др. Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer // *J. Proteome Res.* 2007. Т. 6. № 10. С. 3962–3975.

123. Wilhelm M. и др. Mass-spectrometry-based draft of the human proteome. // Nature. 2014a. Т. 509. № 7502. С. 582–7.
124. Wilhelm M. и др. Mass-spectrometry-based draft of the human proteome. // Nature. 2014b. Т. 509. № 7502. С. 582–7.
125. Wruck W., Peuker M., Regenbrecht C.R. a. Data management strategies for multinational large-scale systems biology projects // Brief. Bioinform. 2014. Т. 15. № 1. С. 65–78.
126. Wu J. и др. Identification and quantification of osteopontin splice variants in the plasma of lung cancer patients using immunoaffinity capture and targeted mass spectrometry. // Biomarkers. 2012. Т. 17. № 2. С. 125–33.
127. Yip Y.L. и др. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase // Hum. Mutat. 2008. Т. 29. № January. С. 361–366.
128. Yocum A.K. и др. Coupled global and targeted proteomics of human embryonic stem cells during induced differentiation. // Mol. Cell. Proteomics. 2008. Т. 7. № 4. С. 750–767.
129. Yost R. a, Enke C.G. Selected Ion Fragmentation with a Tandem Quadrupole Mass Spectrometer // J. Am. Chem. Soc. 1978. Т. 100. № 7. С. 2274–2275.
130. Zgoda V.G. и др. Chromosome 18 Transcriptome Profiling and Targeted Proteome Mapping in Depleted Plasma, Liver Tissue and HepG2 Cells // J Proteome Res. 2013.
131. Zhang H. и др. Methods for peptide and protein quantitation by liquid chromatography-multiple reaction monitoring mass spectrometry. // Mol. Cell. Proteomics. 2011. Т. 10. № 6. С. M110.006593.
132. Zhi W., Wang M., She J.-X. Selected reaction monitoring (SRM) mass spectrometry without isotope labeling can be used for rapid protein quantification. // Rapid Commun. Mass Spectrom. 2011. Т. 25. № 11. С. 1583–8.
133. Zweigenbaum J., Henion J. Bioanalytical high-throughput selected reaction monitoring-LC/MS determination of selected estrogen receptor modulators in human plasma: 2000 samples/day // Anal. Chem. 2000. Т. 72. № 11. С. 2446–2454.
134. Николаев Е.Н. и др. Анализ белков и пептидов методами масс-спектрометрии высокого разрешения // Успехи химии. 2012. Т. 81. № 11. С. 1051–1070.
135. Пономаренко Е.А. и др. Россия в международном проекте “протеом человека”: первые итоги и перспективы // Биомедицинская химия. 2015. Т. 61. № 2. С. 169–175.

136. Рыбина А.В. и др. Простой метод оценки вероятности детекции пептида при масс-спектрометрии с электроспрейной ионизацией // Биомедицинская химия. 2014. Т. 60. № 6. С. 707–712.
137. Скворцов В.С. и др. pIPredict – компьютерная программа для предсказания изоэлектрической точки пептидов и белков // Биомедицинская химия. 2015. Т. 61. № 1. С. 83–91.
138. Чернобровкин А.Л. Анализ масс-спектров масс-спектров пептидных фрагментов для идентификации генетически-детерминированного полиморфизма белков // 2012.

ПРИЛОЖЕНИЕ

Таблица. Протеотипические пептиды контрольной выборки.

№	UniProt AC (Название белка)	Пептидная последовательность	Количество копий пептида в клетке линии HepG2	Клеточная локализация, GO	Молекулярно- биологическа я функция, GO
1.	A6NNZ2 (Тубулин бета 8 подобный белок)	HGCYLTVAAlFR	2.52*10 ⁸	Цитоплазма [GO:0005737]	Активность ГТФаз [GO:0003924]
		LAVNMVPFPR	2.52*10 ⁸		
2.	O60543 (Активатор клеточной смерти CIDEA)	GVMASSLQELISK	3.53*10 ⁵	Цитоплазма [GO:0005737]	Полиаденилир ование [GO:0044822]
		VTFDLR	4.56*10 ⁵		
3.	P18621 (60S рибосомаль ный белок L17)	YSLDPENPTK	4.34*10 ⁶	Цитоплазма [GO:0005737]	Активность кортикотропи нового рецептора[GO :0004978]
		EQIVPKPEEEVAQK	5.10*10 ⁶		
4.	P48594 (Серпин B4)	LVLVNAIYFK	1.47*10 ⁶	Межклеточны й контакт [GO:0005913]	Связывание ионов кальция [GO:0005509]
		VLHFDQVTENTTE K	1.47*10 ⁶		
5.	Q01718 (Рецептор адренкор икотропног о гормона)	NNSDCPR	1.93*10 ⁵	Внеклеточная локализация [GO:0005576]	Структурная активность молекул [GO:0005198]
		NMGYLKPR	1.37*10 ⁵		
6.	Q02413 (Десмоглеи н-1)	YQGTILSIDDNLQR	7.11*10 ²	Цитоплазма [GO:0005737]	
		LADISLGK	8.89*10 ²		
7.	Q02487 (Десмоглеи н-2)	TNEGVLCVVKPLN YEEK	6.15*10 ³		Связывание ионов кальция [GO:0005509]
		SFTILLSNTENQEK	1.30*10 ⁵		
8.	Q13485 (Декапента плегически й белок)	ETPCWIEIHLHR	3.58*10 ⁴	Интегральный компонент клеточной мембраны	связывание Wnt-белков [GO:0017147]
		IYPSAYIK	1.20*10 ⁴		

				[GO:0005887]	
9.	Q15846 (Кластерин - подобный белок-1)	APDHGGLISK	$1.41 \cdot 10^3$	Ядро [GO:0005730]	Связывание повреждённой ДНК [GO:0003684]
		IYQFLFPFHEDNEK	$6.90 \cdot 10^3$		
10	Q16787 (Ламинин субъединица альфа-3)	SFGLPGDMVLLEK	$5.13 \cdot 10^4$	Ядро [GO:0005634]	Активность цистеиновых протеиназ [GO:0004197]
		LVFSIRPR	$2.45 \cdot 10^4$		
11.	Q2VWA4 (Белок ассоциированной с имфомой транслокации)	VPAPHHPHLLLEGR	$1.16 \cdot 10^4$	Клеточная мембрана [GO:0016324]	Связывание микротрубочек [GO:0008017]
		AGLSGLFWPAGR	$1.11 \cdot 10^4$		
12.	Q6PI98 (Субъединица C комплекса INO80)	MVPSEFSTGPVEK	$5.06 \cdot 10^4$	Эндоплазматический ретикулум [GO:0005783]	Активность ГТФаз [GO:0003924]
		FSTIEEFSYIR	$5.03 \cdot 10^4$		
13.	Q6UXH8 (Коллаген - и кальций - связывающий белок)	GSMGPMGPSPDLS HIK	$2.45 \cdot 10^4$	Цитоплазма [GO:0005737]	Полиаденилирование [GO:0044822]
		YPNDTGHEK	$1.37 \cdot 10^4$		
14.	Q8IYS8 (Белок, регулирующий биоориентацию хромосом при клеточном делении)	ISSETSSVESVMGIK	$1.37 \cdot 10^8$	Цитоплазма [GO:0005737]	Активность кортикотропного рецептора [GO:0004978]
		ADVDTKPAYQNLS QK	$1.02 \cdot 10^7$		
15.	Q8J025 (Белок APCDD1)	LPHTYEYEIFK	$3.37 \cdot 10^7$	Клеточная мембрана [GO:0016324]	Связывание ионов кальция [GO:0005509]
		SDEHHPPILPPK	$1.67 \cdot 10^7$		
16.	Q96K83 (Цинковый палец-521)	EFPNQESLLK	$1.30 \cdot 10^2$	Внеклеточная локализация [GO:0005576]	Структурная активность молекул
		FGSPVLGTPK	$3.13 \cdot 10^2$		

					[GO:0005198]
17.	Q99708 (ДНК-эндонуклеаза)	VSSPVFGATSSIK	$2.20 \cdot 10^5$	Базальная мембрана [GO:0005604]	Связывание SMAD [GO:0046332]
		EVQGLQVK	$3.00 \cdot 10^3$		
18.	Q9BXW6 (Оксистеро-л-связывающий белок)	LQELDPATYK	$1.98 \cdot 10^5$	Комплекс I α 80 [GO:0031011]	Связывание ионов кальция [GO:0005509]
		ITMPVIFNEPLSFLQR	$9.21 \cdot 10^6$		
19.	Q9UDY8 (Паракаспа 1)	GTPEETGSYLVS	$1.28 \cdot 10^4$	Кинетохор [GO:0000777]	Связывание Wnt-белков [GO:0017147]
		LSELLDQAPEGR	$3.21 \cdot 10^4$		
20.	Q9Y2V3 (Белок ретикулярного гомеобокса)	DDGILGTFPAER	$9.68 \cdot 10^3$	Ядро [GO:0005634]	Связывание повреждённой ДНК [GO:0003684]
		FPLDEADPR	$1.48 \cdot 10^5$		
21.	Q9Y4B5 (Фактор-1 перекрестного связывания микротрубочек)	AEDENETLR	$8.83 \cdot 10^5$	Комплекс CBM [GO:0032449]	Активность цистеиновых протеиназ [GO:0004197]
		LVEEEANILGR	$6.61 \cdot 10^5$		
22.	Q9Y4J8 (Дистобривин альфа)	GIQYSLNVADR	$6.13 \cdot 10^4$	Ядро [GO:0005634]	Связывание микротрубочек [GO:0008017]
		ELNSEVGSETESNV DSEFAR	$8.43 \cdot 10^4$		
23.	Q9Y5U9 (Белок раннего ответа-1)	NIGWGTDQGIGGF GEEPGIK	$1.12 \cdot 10^4$	Аксон [GO:0030424]	Активность ГТФаз [GO:0003924]
		SQLMNLIR	$2.32 \cdot 10^4$		