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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2015 г.

Работа выполнена в Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича».

Научный руководитель: доктор биологических наук,
член-корреспондент РАН
Лисица Андрей Валерьевич

Официальные оппоненты: Шишкин Сергей Сергеевич
доктор биологических наук, профессор
ФГБУН Институт биохимии имени А. Н. Баха РАН,
лаборатория биомедицинских исследований,
зав. лабораторией

Вяткина Кира Вадимовна
кандидат физико-математических наук,
Санкт-Петербургский Академический университет - научно-образовательный центр нанотехнологий РАН,
лаборатория вычислительной биологии,
старший научный сотрудник.

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт биохимической физики им. Н.М. Эмануэля РАН

Защита состоится «17» декабря 2015 года в 11:00 часов на заседании Диссертационного совета Д 001.010.01 при Федеральном государственном бюджетном научном учреждении «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» (ИБМХ) по адресу: 119121, г. Москва, Погодинская ул., д. 10, стр. 8.

С диссертацией можно ознакомиться в библиотеке и на сайте ИБМХ www.ibmc.msk.ru.

Автореферат разослан «____» 2015 г.

Ученый секретарь Диссертационного совета,
кандидат химических наук  Е.А. Карпова

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

1.1. Актуальность проблемы, цель и задачи

Повышение эффективности методов исследования биологических молекул за последние 20 лет сделало возможным выполнение широкомасштабных проектов, таких как «Геном человека» [Venter и др., 2001], «Протеом человека» [Orchard, Ping, 2009]. Масштабность подразумевает, что в результате не только появляется информация об исследуемом объекте, но и возникает массив сведений, формирующий направление будущих исследований [Pearsons, 1991]. Результаты широкомасштабных проектов требуют интерпретации и уточнения.

В 2010 был дан старт международному хромосомоцентричному 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 исследования, отработанные на одной хромосоме, переносимы и на другие хромосомы.

Россия, как страна-участница международного проекта «Протеом человека», провела масс-спектрометрические измерения для белков хромосомы 18 человека [Ponomarenko и др., 2014]. Для измерений использовали направленный масс-спектрометрический метод мониторинга реакций диссоциации протеолитических пептидов белков, так называемых диссоциативных переходов (метод МДП). Метод МДП является методом направленной масс-спектрометрии, то есть основан на регистрации масс-спектрометром заранее заданных исследователем отношений массы к заряду целевого соединения (например, пептида) и его фрагментов. Метод основан на количественной идентификации пептидов по масс-спектрам ионов-фрагментов полученных в результате столкновений родительских ионов пептидов с нейтральными молекулами буферного газа (как правило, гелий, или азот). При этом, в смеси протеолитических пептидов белка определяются пептиды, дающие наиболее интенсивные сигналы в масс-спектрах. В первую очередь, это пептиды, образующиеся в наибольшей концентрации в процессе ферментативного гидролиза соответствующего белка, и с наиболее высокой эффективностью ионизации в источнике ионизации масс-спектрометра. Такие пептиды называются

протеотипическими и именно они являются основным объектом количественных измерений методом МДП.

На сегодняшний день, метод МДП является одним из наиболее чувствительных методов протеомики. Он позволяет проводить измерения единичных молекул белков в клетке [Picotti и др., 2009], а также белков в концентрациях около нг/мл в плазме крови [Keshishian и др., 2009; Kuzyk и др., 2009]. Несмотря на высокую чувствительность метода мониторинга диссоциативных переходов, достоверность измерения сложно оценить из-за интерференции пептидных ионов в биологической матрице [Bao и др., 2013; Sherman и др., 2009; Sherman, Molloy, Burlingame, 2012]. Интерференцией называют регистрацию масс-спектрометрического сигнала соответствует не одному пептиду, а нескольким неизвестным компонентам сложной смеси. До сих пор в практике метода МДП отсутствуют общепринятые средства статистической оценки достоверности детекции пептида. При этом существующие алгоритмы требуют изменения протокола проведения эксперимента, например, настройки прибора на масс-зарядные характеристики заведомо несуществующих соединений или добавления в биологический материал стандартизирующих растворов синтетических пептидов. Разработка некоторых алгоритмов проводилась на обучающих выборках, состоящих из высококопийных белков [MacLean и др., 2010; Reiter и др., 2011], тогда как при хромосомоцентричном подходе проблема заключается в выявлении низкокопийных белков.

Для решения этой проблемы в настоящей работе были собраны и приведены к унифицированному виду данные о результатах масс-спектрометрической детекции белков хромосомы 18 человека. Массив данных был обработан с помощью системы фильтров, основанных на анализе воспроизводимости регистрируемых в ходе масс-спектрометрического эксперимента параметров. В результате была получена оценка уровня технических ошибок, ограничивающих использование масс-спектрометрического метода МДП для исследования молекулярных процессов.

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.

1.2. Положения, выносимые на защиту

1. Разработанный алгоритм обработки результатов хромосомоцентричных масс-спектрометрических измерений позволяет выявлять протеотипические пептиды, обеспечивающие количественный анализ белков человека в сложных биологических матрицах.

2. Оценка воспроизводимости результатов количественных измерений протеотипических пептидов позволяет систематизировать результаты широкомасштабных протеомных исследований.

1.3. Научная новизна и практическая значимость

В работе проведен анализ протеомных данных, полученных направленным масс-спектрометрическим методом в рамках хромосомоцентричного исследования. Исследование белков одной хромосомы в биоматериале позволяет выработать формальные критерии для оценки качества результатов направленных масс-спектрометрических измерений. Разработанный подход послужит методической основой стандартизации результатов, полученных методом мониторинга диссоциативных переходов.

Практическая значимость работы заключается в формировании методики для отбора протеотипических пептидов для количественного анализа белков. На основании обработанных данных сформирована база данных масс-спектрометрических измерений, содержащая результаты измерений для 2247 протеотипических пептидов белков, кодируемых генами хромосомы 18. В рамках базы данных выделены 46 пептидов, измерения которых характеризуются наибольшей воспроизводимостью.

Предлагаемый в работе алгоритм позволяет выявить из противоречивого информационного массива данные, которые обладают достаточной воспроизводимостью для характеристики протеомного состава биологического материала (клеточной линии HepG2, плазмы крови человека). Разработанный алгоритм позволяет выявлять биологически значимые явления, в частности, исследовать трансляцию сплайс-опосредованных вариантов белков. Анализ данных, накопленных для белков, кодируемых генами одной хромосомы, показал, что значительная часть измерений искажается вследствие влияния сложной биологической матрицы. Из этого следует, что в различных типах биоматериала необходимо проводить подбор протеотипических пептидов и фрагментных ионов для масс-спектрометрического анализа.

1.4. Личный вклад автора

1. Создание схемы алгоритма анализа массива экспериментальных данных.
2. Разработка структуры данных и интерфейса базы данных.

3. Обработка массива экспериментальных данных с помощью разработанного алгоритма.
4. Осуществление аннотации сформированной выборки пептидов.

1.5. Апробация работы

Основные положения диссертационной работы были представлены в виде постерного доклада на 13-м Ежегодном всемирном конгрессе Международной организации «Протеом человека» (HUPO 13-th Annual World Congress, Мадрид, 2014). Постерные сообщения представлялись также на научном конгрессе «Протеомный форум» (Proteomic forum, Берлин, 2013); на XX Российском национальном конгрессе «Человек и лекарство» (Москва, 2013); на конгрессе Федерации европейских биохимических обществ 2013 «Биологические механизмы» (FEBS congress, St.Petersburg, 2013); на 12-м ежегодном конгрессе Международной организации «Протеом человека» (HUPO 12-th Annual World Congress, Йокогама, 2013).

1.5. Публикации

По теме диссертационной работы опубликовано 11 работ, из которых 6 статей в международных рецензируемых изданиях и 5 публикаций в трудах конференций.

1.6. Объем и структура диссертации

Диссертационная работа изложена на 128 страницах машинописного текста; содержит 7 таблиц и 21 рисунок. Состоит из глав: «Введение», «Обзор литературы», «Материалы и методы», «Результаты и обсуждение», «Заключение», «Выводы», «Список литературы»; включает 1 приложение.

2. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

2.1. Исходные 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]. Массив данных представлял собой набор хроматограмм в проприетарном формате (Agilent, США) и включал экспериментальные данные, полученные для двух биологических проб: клеточной линии HepG2 и белкового экстракта плазмы 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 отчетных 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 xslx с конфигурационными настройками хромато-масс-спектрометрической системы Agilent 6490 TripleQuadrupole 6490 (США), в том числе перечень масс-спектральных характеристик фрагментных ионов протеотипических пептидов, использованных для проведения измерений.

Общий объем использованных для выполнения работы исходных данных составил 38 ГБ. Данные были сгенерированы с использованием хромато-масс-спектрометрической станции под управлением ПО MassHunter Data Acquisition B.04.

2.2. Обработка данных, полученных методом мониторинга диссоциативных переходов

Исходные данные обрабатывали программным обеспечением масс-спектрометра с масс-анализатором типа «тройной квадруполь» MassHunter Qualitative Analysis B.4.0.

Результат обработки сохраняли в виде электронного отчета, используя следующие настройки: включение в состав отчета списка пептидных ионов и соответствующих им хроматограмм, а также ионов – фрагментов, характеристик и изображений их хроматографических пиков. Электронный отчет («CompoundReport») представлял собой папку, содержащую графические файлы с изображениями хроматограмм фрагментных ионов в формате .emf, а также файл .xml, содержащий характеристики аннотированных оператором групп пиков.

Для проверки соответствия исходных данных разработанной модели нами была разработана вспомогательная программа «CompoundReportConverter.pl». Программно анализировался файл карты фрагментации (.xls или .xlsx), который содержал настройки, использовавшиеся при выполнении измерений, а также электронные отчеты (/Compound Report). При выявлении в составе этих файлов противоречивых сведений или отсутствия необходимых для дальнейшей работы данных файлы считали непригодными для анализа.

2.3. Программная реализация базы данных и система запросов

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). Доступ к базе данных осуществляется посредством интернет-обозревателя по адресу http://www.pikb18.ru/.

Система запросов к информационному ресурсу была реализована в виде комплекса подпрограмм, написанных на языке программирования Perl. Запросы *q(x)* обеспечивали извлечение информации в соответствии с взаимоотношениями между объектами предметной области, установленными моделью данных:

$$q(Exp) = (\{Prot1, Prot2 \dots ProtN\} \mid \{Pept1, Pept2 \dots PeptN\} \mid \{ Sample1 \mid Sample2\}) \quad (1)$$

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

$$q(Pept) = (Prot \mid (FragMap, \{Run1, Run2 \dots RunN\})), \text{ где} \quad (3)$$
$$Run = \{tr1, tr2 \dots trN\}; tr = [ApexRt, Intesity, SNR]$$

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; *Sample*, биологический материал. Результатом выполнения запросов (1)–(3) являлась выборка экспериментов и технических повторов, соответствующих одному или нескольким критериям, сформулированных в запросе.

2.4. Оценка результатов масс-спектрометрических измерений

Оценку технической воспроизводимости проводили для определения погрешностей в режиме работы прибора, в пробоподготовке, а также случайных ошибок при отборе протеотипических пептидов и их фрагментных ионов. Степень технической вариабельности использовали, чтобы выявить измерения с наибольшей воспроизводимостью и, таким образом, определить, является ли регистрируемый сигнал биологически значимым. Если разница в концентрациях пептидов одного белка (парных пептидов) превышала техническую воспроизводимость, то анализировали биологически обусловленные причины, например, альтернативный сплайсинг.

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

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

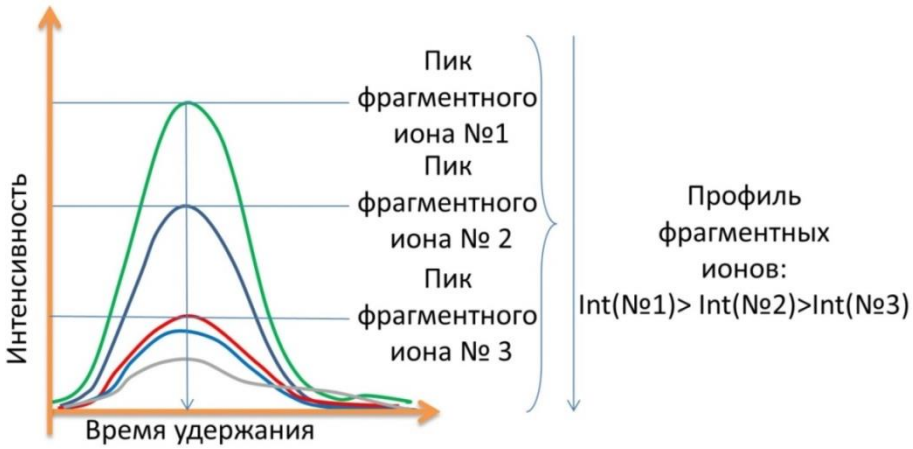

Рисунок 1. *Группа хроматографических пиков фрагментных ионов, возникающих в результате индуцированной диссоциации пептидного иона в столкновительной ячейке масс-спектрометра. Int – интенсивность пика, оцениваемая как его высота или как площадь*

Профиль фрагментных ионов пептида оценивали как показано на рисунке 1. 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. В повторном измерении анализировали, сохранился ли порядок следования ионов с учетом допущения на приборную погрешность в пределах 10% от интенсивности пика. Коэффициент вариации рассчитывали как число совпавших по порядку следования ионов N_1 к их общему числу N:

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

Оценку воспроизводимости проводили после сегментирования данных. Сегментирование заключалось в распределении белков по четырем 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, трех или более порядков (группы 2–4, соответственно). Содержание целевого белка выражали как количество копий молекулы уникального пептида на клетку линии HepG2.

2.5. Биоинформатические ресурсы

Для получения данных о регистрации белков хромосомы 18 на транскриптомном и протеомном уровнях, локализации белков в клеточных компонентах, а также вовлеченности в молекулярно-биологические процессы и механизмы развития заболеваний, использовали БД UniProt[1].

Для каждого белка хромосомы 18 из БД UniProt загружали сведения о его детекции экспериментальными протеомными методами или данные об экспрессии белок-кодирующего гена.

Аминокислотные последовательности вариантов, образованных в результате альтернативного сплайсинга, также загружали из UniProt[2] и подвергали виртуальному трипсинолизу. Уникальность последовательности каждого детектированного протеотипического пептида белков хромосомы 18 проверяли по отношению к каноническим формам и сплайс-опосредованным вариантам белков в масштабе всего протеома.

Базу данных Plasma Proteome Database (PPD)[3] [Nanjappa 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 человека, список белков, детектированных методом мониторинга диссоциативных переходов, а также доступные из литературных источников сведения об измеренных концентрациях белков в плазме здоровых людей.

Информацию о концентрации белков в клетках линии HepG2 получали из базы данных MaxQB, содержащей результаты панорамных протеомных исследований 11 клеточных линий [Schaab и др., 2012]. Всего использовали данные

---

[1] http://www.uniprot.org/uniprot/

[2] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot_varsplic.fasta.gz

[3] http://www.plasmaproteomedatabase.org/

о содержании 80 белков, кодируемых генами хромосомы 18 человека в клетках 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 RESTful-сервиса[4]. Используя в качестве поискового запроса идентификаторы UniProt AC, 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 в 166 экспериментах.

2.6. Специализированное программное обеспечение

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

Сведения о клеточной локализации белков и их роли в молекулярно-биологических процессах получали с использованием программного обеспечения GOrilla [Eden и др., 2009].

Контроль соответствия детектированных пептидных последовательностей белкам хромосомы 18, а также контроль соответствия масс-зарядных характеристик пептидных ионов и ионов-фрагментов, использованных для настройки прибора, осуществляли с помощью ПО Skyline 2.6[5] [MacLean и др., 2010].

Расчет теоретического времени удержания пептида проводили с использованием ПО Theoretical Chromatograph [Gorshkov и др., 2006] для оценки контрольной выборки протеотипических пептидов. Данные об интерференции пептидных ионов получали с помощью ПО SRM Collider, версия 1.4 [Röst, Malmström, Aebersold, 2012].

[4] http://www.ebi.ac.uk/pride/ws/archive/
[5] https://skyline.gs.washington.edu/labkey/project/home/software/Skyline/begin.view

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

3.1. Модель данных для описания результатов направленного масс-спектрометрического эксперимента

Разработанная модель данных включает информационные объекты, описывающие процесс и результаты детекции белков методом мониторинга диссоциативных переходов. Представленная на рисунке 2 модель включает следующие объекты масс-спектрометрического эксперимента: «белок», «пептид», «фрагмент», «эксперимент», «технический повтор» на различных этапах технологического процесса. 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 модель данных позволяет оперировать информационными объектами с целью оценки влияния допущений на итоги протеомного исследования.

Объекты модели связаны друг с другом: «один ко многим» и «многие ко многим». Пептид может относиться только к одному белку (требование уникальности), при этом один белок может быть связан с несколькими пептидами. Пептид, как объект информационной модели, обладает свойством *протеотипичности* – то есть данный пептид достоверно идентифицируется масс-спектрометрически. Понятие протеотипичности пептида включает «уникальность»: аминокислотная последовательность пептида соответствует канонической (наиболее часто встречающейся, согласно 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*.

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 могут относиться либо к разным, либо к одинаковым

фрагментам/пептидам/белкам. Если связь белков/пептидов с экспериментами/техническими повторами является результатом планирования эксперимента, то различия между техническими повторами одного эксперимента отражают вариабельность на техническом (методическом) и биологическом уровне. Это ключевое свойство модели данных позволяет посредством системы запросов отобрать белки, измерения которых в наименьшей степени 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.

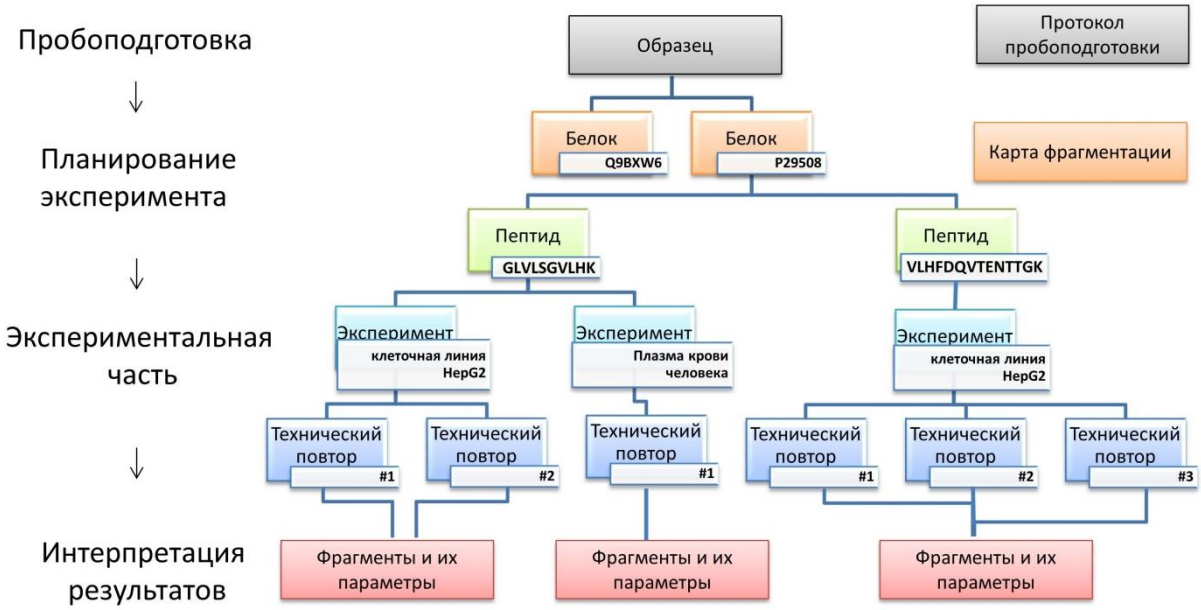

Рисунок 2. *Информационная модель данных протеомных исследований, основанных на масс-спектрометрическом методе МДП.*

Модель данных, предложенная в настоящей работе, потребовала разработки системы информационных запросов. Запросы были реализованы в соответствии с формулами (1) – (3), приведенными в разделе 2.5.

Система запросов позволила охарактеризовать результаты загрузки экспериментов в базу данных протеотипических пептидов. Результатом загрузки в базе данных стали сведения о 1689 экспериментах, выполненных масс-

спектрометрическим методом мониторинга диссоциативных переходов для исследования 275 белках, закодированных на хромосоме 18 человека, в том числе были данные о результатах измерений в гидролизатах клеточной линии и плазме крови. В совокупности, для всех экспериментов была загружена информация о 2247 протеотипических пептидах. Более чем для половины пептидов было детектировано по 5 диссоциативных переходов, большинство из которых были измерены с соотношением сигнал-шум, превышающим 10.

3.2. Оценка результатов детекции белков и пептидов в биологических образцах

Данные технических повторов использовали для оценки воспроизводимости результатов экспериментов. Полагали, что чем более воспроизводим результат при разных условиях постановки эксперимента, тем выше правдоподобие гипотезы о детектировании пептидов и, следовательно, тем точнее количественная оценка, выполненная масс-спектрометрическим методом МДП.

На рисунке 3 показаны варианты распределений коэффициента вариации для различных параметров, регистрируемых в ходе эксперимента, выполняемого методом МДП. Для времен удержания на хроматографической колонке, как показано на рисунке 3а, характерен низкий коэффициент вариации (менее 1%). Это объясняется тем, что в используемом для обработки первичных данных ПО Mass Hunter заложена функция определения группы пиков, соответствующих одному соединению, по совпадению времен удержания. То есть, если фрагменты, возникающие в результате диссоциативных переходов, действительно принадлежат одному пептиду, то они должны регистрироваться в одно и то же время. Другая группа параметров (см. рис. 3б и рис. 3в) характеризуется распределением коэффициента вариации, имеющим выраженный максимум. Например, коэффициент вариации интенсивности в наибольшем количестве экспериментов составляет от 50 до 150 процентов (см. рис. 3в). Характеристики, распределение коэффициента вариации которых имеет максимум, а соотношение среднего значения и медианы близко к единице, использовались для разработки каскада фильтров с целью отбора пептидов с большей правдоподобностью детекции и меньшей погрешностью в количественной оценке.

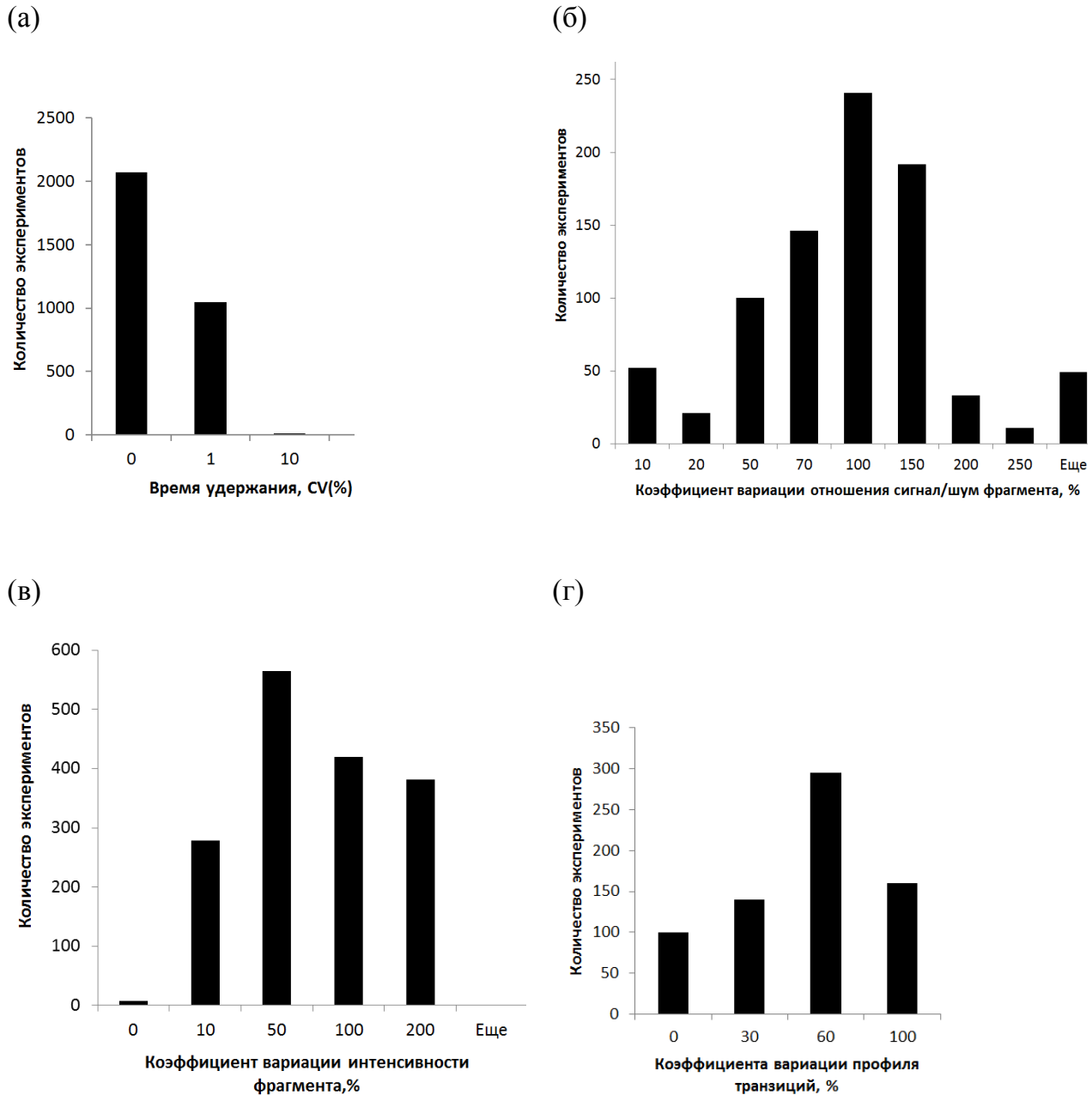

Рисунок 3. *Распределение коэффициентов вариации характеристик диссоциативных переходов протеотипических пептидов:(а) время удержания; (б) соотношение сигнал/шум фрагментных ионов; (в) интенсивность фрагментных ионов пептидов; (г) профилей фрагментных ионов пептидов.*

С помощью анализа технических повторов отбирали группу пептидов, результаты количественного измерения которых были наиболее воспроизводимы. Далее проводили количественные оценки относительных концентраций парных протеотипических пептидов. Если пептиды действительно относятся к одному

белку, то они должны находиться в эквимолярных концентрациях, следовательно, их концентрации в биоматериале не должны различаться существенным образом.

Исходя из вышеизложенного, был разработан каскад фильтров для отбора пептидов с воспроизводимыми значениями интенсивности и совпадающими профилями диссоциативных переходов. Из массива данных, который изначально насчитывал 2247 пептидов для 275 белков, более полутора тысяч пептидов было отвергнуто из-за низкой интенсивности пиков диссоциативных переходов. Применение следующего фильтра, основанного на сопоставлении профилей транзиций (пар пептидного иона и иона фрагмента), сократила контрольную выборку до 86 пептидов (65 белков).

Среди оставшихся белков только для 23-х имелись данные о результатах измерений двух пептидов, относящихся к одному белку. Контрольная выборка была распределена на четыре группы. В первую группу вошли 12 белков, для которых результаты измерений парных пептидов различались не более, чем в два раза. Группа №2 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 относительных концентраций, полученных для разных пептидов данного белка и превышающими три порядка величины, были отнесены 4 белка и 8 пептидов. Примеры хроматограмм пептидов разных групп приведены на рисунке 4. Хроматограммы пептидов из первой и второй группы характеризуются совпадением вершин хроматографических пиков ионов-фрагментов, а также сходством их формы. Видно, что на хроматограммы пептидов из третьей и четвертой групп хроматографические пики менее выражены и отличаются по форме. Интенсивность пиков всех пептидов, представленных на рисунке 4, находится на примерно одном уровне.

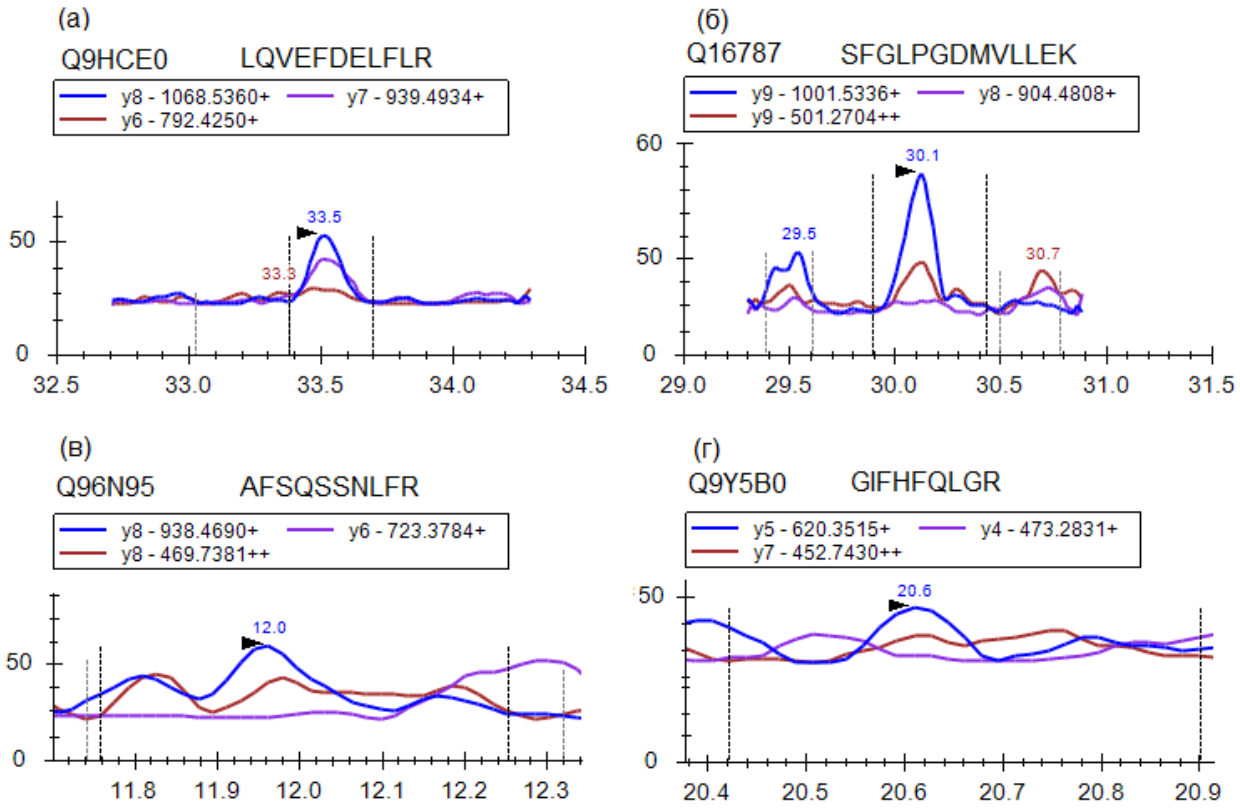

Рисунок 4. *Хроматограммы пептидов из (а) первой, (б) второй, (в) третьей и (г) четвертой групп.*

3.3. Анализ выборки протеотипических пептидов и соответствующих им белков

Контрольную выборку из 23 белков охарактеризовали с использованием активно применяемых в протеомике информационных ресурсов UniProt и GeneOntology (GO).

На рисунке 5 показаны свойства выборки белков хромосомы 18. Например, в хромосоме 18 доля транслируемых в белки генов превышает 65% , причем и в других хромосомах данный показатель находится примерно на таком же уровне [Lane и др., 2014]. Как следует из рисунка 5а, соотношение в контрольной выборке остается неизменным: отобранные с использованием каскада фильтров белки превалируют в категории «протеомный уровень». Аналогичная ситуация наблюдается на рисунке 5б, где приведены данные по количеству копий пептидов в единичной клетке HepG2. Видно, что в контрольной выборке, как и во всей хромосоме 18, максимальное количество пептидов представлено в количестве 10 тысяч копий на клетку HepG2.

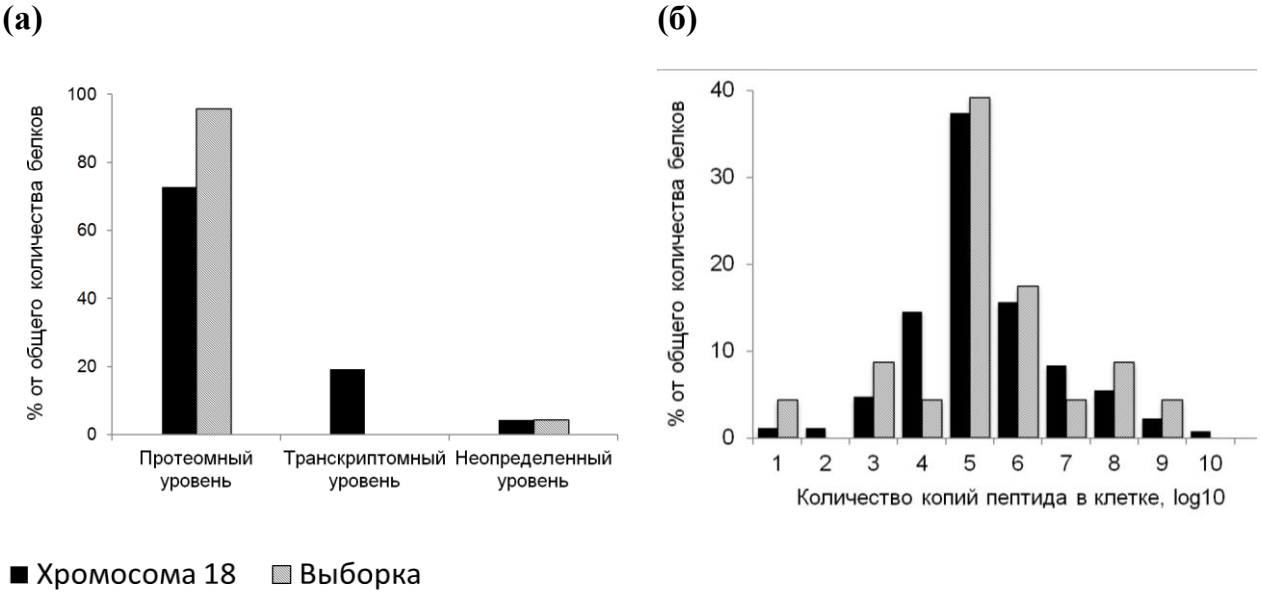

Рисунок 5. *Сопоставление хромосомы 18 (276 белков) и контрольной выборки из 23-х белков, полученной в результате анализа свойств протеотипических пептидов: (а) уровень реализации генетической информации согласно ресурсу UniProt; (б) распределение количества копий пептидов в расчете на единичную клетку линии HepG2.*

Сопоставляя онтологию генов (GO) хромосомы 18 с аннотацией всего генома, с вероятностью ошибки менее 10^{-4}, получили, что гены хромосомы 18 обеспечивают формирование межклеточных десмосомальных контактов. Для выборки из 23 белков, сформированной на основе оценки правдоподобия масс-спектрометрических результатов, получено, что 13% генов принадлежат к категории «межклеточный» контакт. Это означает, что в плане клеточных функций фокусированная выборка *конгруентна* исследуемой хромосоме.

Применение информационных фильтров позволило получить несмещенную выборку, как в отношении сведений об экспрессии генов, так и в отношении количества копий белковых продуктов в единичной клетке (см. рис. 5). Контрольную выборку с масс-спектрометрическими данными, полученными направленным методом МДП в плазме крови человека [Ponomarenko и др., 2014; Zgoda и др., 2013] и панорамным (MS/MS) методом в клеточной линии HepG2 [Schaab и др., 2012]. Получили, что 15 из 23 белков в составе контрольной 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]. Корреляция повышалась при сопоставлении в пределах контрольной выборки (R^2 = 0.59), однако достоверность корреляционной связи была низкая, так как основное количество точек сгруппировано в середине диаграммы.

Низкий уровень корреляции, вероятно, связан с использованием разных пептидов при детектировании одних и тех же белков. Например, белок P18621 был детектирован направленным методом МДП с использованием пептидов EQIVPKPEEEVAQK и YSLDPENPTK, тогда как при использовании панорамного метода для расчета количественного содержания использовали другие пептиды: GIDVDSIVIEHIQVNK, QWGWTQGR, SAEFIIHMIK.

Дальнейший анализ выборки белков проводили с использованием ресурса Plasma Proteome DB [Nanjappa и др., 2014]. В этом ресурсе обнаружились сведения только для трех белков хромосомы 18. Концентрации белков десмоглеина-1 (Q02413), десмоколлина-2 (Q02487) и серпина B4 (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%.

3.4. Выявление сплайс-опосредованных вариантов

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

Сравнение детектированных пептидных последовательностей, вошедших в состав контрольной выборки, на последовательности сплайс-вариантов белков показало, что некоторые протеотипические пептиды входят в состав сплайс-форм 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.

Например, пептид LAVNMVPFPR бета-тубулина входит в состав сплайс-формы третьей субъединицы бета-тубулина. Разница с концентрацией парного

пептида HGCYLTVAAIFR составляет два порядка. Показательным является пептид LVLVNAIYFK серпина B4, соответствующий сплайс-формам трех белков человека: интерлейкин связывающего фактора, серпина B5 и серпина B8 P50452-3. Разница его концентрации с парным пептидом VLHFDQVTENTTEK составила более шести порядков.

Наблюдались случаи, когда протеотипические пептиды картировались на разные сплайс-опосредованные варианты одного белка, при этом концентрации парных пептидов были в пределах одного порядка. Например, пептиды GTPEETGSYLVSK и LSELLDQAPEGR транслоцируемого при лимфомах белка (Q9UDY8) соответствуют как его канонической (мастерной) форме, так и его сплайс-опосредованному варианту Q9UDY8-2. Кроме того, пептид GTPEETGSYLVSK картировался еще на один сплайс-опосредованный вариант этого же белка (Q9UDY8-3).

В составе контрольной выборки было 14 белков, пептиды которых соответствовали исключительно канонической форме целевого белка, то есть не совпадали с известными в масштабах протеома сплайс-опосредованными вариантами. Для таких белков измеренные концентрации парных пептидов отличались не более, чем в два раза, что в количественной масс-спектрометрии считается в пределах приборной погрешности.

Рассмотренные примеры показывают принцип применения хромосомоцентричной базы данных протеотипических пептидов. Информационная модель 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 контрольной выборки белков. Выборка, будучи ограниченной по масштабу, полностью наследует информационные свойства исходной хромосомы. Анализ контрольной выборки позволяет планировать дальнейшие исследования, например, основанные на масс-спектрометрическом измерении функционально значимых форм белков.

Таблица 2. Сплайс-опосредованные варианты в составе контрольной выборки белков, полученной в результате каскадной фильтрации и сегментации базы данных протеотипических пептидов хромосомы 18 человека.

Название (UniProt AC)**	**Аминокислотная последовательность пептида**	**Сплайс-опосредованные варианты (UniProt AC)**	**Δ***
Бета-тубулин 8 (A6NNZ2)	HGCYLTVAAIFR	-	10^2
	LAVNMVPFPR	Третья субъединица бета-тубулина (Q40665-2)	
Оксистерол-связывающий белок (Q9BXW6)	LQELDPATYK	Анкирин домен-содержащий белок (Ankyrin repeat domain-containing protein Q6P6B7-2)	10^3
	NDFSIWSILR	Сплайс-опосредованные варианты: (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 протеотипических пептидов, относящиеся к одному белку;*

*** UniProt AC – идентификатор белка в базе данных UniProt.*

4. 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.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Чернобровкин А.Л., Митькевич В.А., Попов И.А., Индейкина М.И.,Ильгисонис Е.В., Лисица А.В., Арчаков А.И. Выявление одноаминокислотных полиморфизмов белков в масс-спектрах пептидных фрагментов // Доклады Академии Наук. 2011. Т. 437(4). С: 561-564.

2. Пономаренко Е.А., Ильгисонис Е.В., Лисица А.В. Технологии знаний в протеомике // Биоорганическая химия. 2011. Т. 37. № 2. С. 190–198.

3. Ponomarenko E.A.; Poverennaya E.V.; Pyatnitskiy M.A.; Lisitsa A.V.; Moshkovskii S.A.; Ilgisonis E.V.; Chernobrovkin A.L.; Archakov A.I. Comparative ranking of human chromosomes based on post-genomic data // OMICS : A journal of integrative biology. 2012. V. 16 (1), P. 604–11.

4. Zgoda VG, Kopylov AT, Lisitsa AV, Ponomarenko EA, Poverennaya EV, Radko SP, Khmeleva SA, Kurbatov LK, Filimonov AD, Bogolyubova NA, Ilgisonis EV *et al*. Chromosome 18 transcriptome profiling and targeted proteome mapping in depleted plasma, liver tissue and HepG2 cells // J. Proteome Res. 2013. V 12(1). P. 123-134.

5. Ponomarenko E.A., Kopylov A.T., Lisitsa A. V., Radko S.P., Kiseleva Y.Y., Kurbatov L.K., Ptitsyn K.G., Tikhonova O. V., Moisa A.A., Novikova S.E., Poverennaya E. V., Ilgisonis E. V. *et al.* 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.

6. Мирошниченко Ю.В., Петушкова Н.А., Москалева Н.Е., Теряева Н.Б., Згода В.Г., Ильгисонис Е.В., Беляев А.Ю. Оценка возможности использования масс-спектрометрической панели пептидов PlasmaDeepDive™ в клинической диагностике// Биомедицинская химия, 2015, Т.61(2), С.272-278.

7. Ekaterina V. Ilgisonis, Arthur T. Kopylov, Elena A. Ponomarenko, Andrey V. Lisitsa. Approximative statistical approach for absolute quantification of the Human Chromosome 18 proteins // In: Proceedings the HUPO 12th Annual World Congress. Yokohama. 2013. P.84.

8. Ekaterina Ilgisonis, Elena Ponomarenko, Andrey Lisitsa. Database for collecting, sharing and comparative analysis of Selected Reaction Monitoring (SRM) Transitions // In: Proceedings the Proteomic Forum 2013. Berlin. 2013. P.138.

9. Ильгисонис Е.В., Информационная система для хранения результатов идентификации белков методом мониторинга множественных реакций // Сборник трудов XX Российского национального конгресса «Человек и лекарство». Москва, 2013, С.346.

10. Ильгисонис Е.В., Копылов А.Т., Лисица А. В., Оценка качества масс - спектрометрических измерений, Сборник трудов XXI Российского национального конгресса «Человек и лекарство», Москва, 2014, С.252-253.

11. Ilgisonis EV, Kopylov AT, Lisitsa AV. Non-redundant peptides database for chr18 encoded proteins quantification // In: Proceedings the HUPO 13th Annual World Congress, Madrid. 2014. P.84.