Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)»

На правах рукописи

Токарева Алиса Олеговна

Оптимизация обработки масс-спектрометрических данных для выявления особенностей липидома клинических образцов

02.00.02 – Аналитическая химия

АВТОРЕФЕРАТ

Диссертации на соискание учёной степени

кандидата физико-математических наук

Москва 2021

Работа выполнена в Институте энергетических проблем химической физики им. В.Л. Тальрозе Федерального государственного бюджетного учреждения науки Федерального исследовательского центра химической физики им. Н.Н. Семенова Российской академии наук.

Научный руководитель: Кандидат физико-математических наук, Кононихин Алексей Сергеевич

Ведущая организация: Федеральное государственное бюджетное научное учреждение «Научно-исследовательский институт биомедицинской химии имени В. Н. Ореховича»

Защита состоится ____ 2021 г. в __:00 часов на заседании диссертационного совета ФЭФМ.02.00.02.001 на базе МФТИ по адресу: 141701, Московская обл., г. Долгопрудный, Институтский пер., д. 9.

С диссертацией можно ознакомиться в научно-технической библиотеке МФТИ и на сайте https://mipt.ru/education/post-graduate/soiskateli-fiziko-matematicheskie-nauki.php

Работа представлена 8 октября 2021 г. в Аттестационную комиссию федерального государственного автономного образовательного учреждения высшего образования “Московский физико-технический институт (национальный исследовательский университет)” для рассмотрения советом по защите диссертаций на соискание ученой степени кандидата наук, доктора наук в соответствии с п. 3.1 ст. 4 Федерального закона “О науке и государственной научно-технической политике”.

ОБЩАЯ ХАРАКТЕРИСТИКА ДИССЕРТАЦИИ

Актуальность темы исследования: Липидомика, как область знаний о липидах, привлекает всё больше внимания вследствие накопления информации о нарушениях метаболизма липидов при различных заболеваниях. Так, на данный момент уже установлены связь между синтезом эйкозаноидов и воспалительными процессами и инфекционными заболеваниями, ингибирование синтеза липопротеинов при диабете, связь между нарушениями в синтезе холестерина и болезнями мозга (болезни Аддингтона, синдрома Альцгеймера, синдрома Паркинсона), связь метаболизма гликосфинголипидов и аутоиммунных заболеваний, а также метаболизма жирных кислот и степени агрессивности рака. Липидомика на сегодняшний день широко задействована в медицине для получения молекулярных профилей тканей, поиске биомаркеров и анализе метаболических путей при различных заболеваниях. Развитие мягких методов ионизации биомакромолекул, таких как электрораспыление (ЭРИ) и матрично-лазерная десорбция с ионизацией (МАЛДИ) привело к тому, что масс-спектрометрия стала широко применяться в липидомных исследованиях. Ввиду разнородности исследуемых образцов биологических тканей, а также широкого профиля заболеваний, потенциально связанных с метаболизмом липидов, разработка методов анализа данных о липидном профиле, полученном с использованием масс-спектрометрии, для клинических целей является актуальной исследовательской задачей.

Степень разработанности темы исследования: На сегодняшний день мало внимания уделено решению проблем деконволюции спектров, получаемых с использованием масс-спектрометрии с прямой ионизацией. При сравнительном исследовании методов нормализации наибольшее внимание уделяется методам с использованием внутреннего стандарта или контроля качества, которые удорожают или удлиняют исследования, игнорируя оценку эффективности методов на основе исследуемых данных. Также при сравнительном анализе

методов селекции переменных при анализе масс-спектрометрических данных игнорируется адекватность применимости выбора переменных для логистической регрессии, которая является основным инструментом в построении диагностических моделей.

Цель исследования: Разработка методик, применимых при анализе данных о липидном профиле биоматериалов различного происхождения, полученных с использованием масс-спектрометрии высокого разрешения для клинических исследований и поиска биомаркеров заболеваний различного генезиса.

Задачи исследования:

1. Повышение эффективности анализа данных, полученных масс-спектрометрическим анализом липидного профиля тканей с использованием прямой ионизации.
2. Сравнительное исследование методов нормализации данных на основе исследуемых образцов, полученных с использованием высокоэффективной жидкостной хромато-масс-спектрометрии.
3. Оптимизация методов поиска биомаркеров в масс-спектрометрических данных, полученных масс-спектрометрическим анализом липидного профиля тканей для дискриминационных и диагностических моделей.

Научная новизна: Разработан метод деконволюции пиков от липидов с одинаковой изобарической массой на основе спектров фрагментации без добавления внутренних стандартов. Автомасштабирование определено как метод, эффективно убирающий межпартийные различия в масс-хроматограммах при сохранении различий в профилях, относящихся к разным клиническим группам. Показана эффективность моноклассовых моделей для построения дискриминационных моделей. Разработан двухстадийный метод поиска

переменных для построения диагностических моделей на основе логистической регрессии. Показана более высокая устойчивость селекции переменных по информационному критерию Акаике, чем с использованием ЛАССО, для построения логистических регрессий.

Практическая значимость: Разработан подход к деконволюции спектра, получаемого с использованием прямой масс-спектрометрии без дополнительной обработки образца. Показана эффективность методов нормализации данных для удаления межпартийных изменений на основе исследуемых данных, что позволяет уменьшить материальные и временные затраты на исследования. Дискриминационные модели на основе выбранного класса липидов упрощают потенциальный целевой анализ и могут быть использованы для дальнейшего исследования патогенеза заболевания. Двухстадийный метод выбора переменных для логистической регрессии может быть использован для широкого круга омиксовых исследований.

Методология и методы исследования: Образцы биоматериала, использованные в работе, были получены от пациенток ФГБУ «НМИЦ АГП им. Кулакова» в ходе клинических исследований, связанных с быстрой и малоинвазивной диагностикой границ опухоли молочной железы, поиском маркеров метастазирования в регионарные узлы при раке молочной железы, сравнительном исследовании изменении метаболизма в эндометриоидных тканях при эндометриозе и миометрии в секреторную и пролиферативную фазу, диагностике рака шейки матки по молекулярному профилю биопсийного материала, диагностике миомы по плазме крови, диагностике эндометриоза по плазме крови, диагностике преэклампсии и задержки внутриутробного развития по плазме крови. Исследования одобрены экспертной комиссией ФГБУ «НМИЦ АГП им. В.И. Кулакова» Минздрава России по вопросам медицинской этики. Все

пациентки подписали добровольное информированное согласие на участие в клинических исследованиях.

Положения, выносимые на защиту:

1. Алгоритм, позволяющий проводить деконволюции сигнала от изобарических ионов.
2. Метод нормализации на основе автомасштабирования для редуцирования отклонений между разными партиями в ходе крупномасштабного хромато-масс-спектрометрического анализа клинических образцов $I^*_{p,s,b} = \frac{I_{p,s,b} - <I_{p,b}>}{sd(I_{p,b})} * sd(I_p) + < I_p >$.
3. Дискриминационная модель на основе отдельных классов липидов пригодных для классификации образцов и для оценки вклада липидов в развитие заболевания.
4. Двухстадийный метод выбора переменных с использованием значений проекций переменной и информационного критерия Акаике при поиске молекулярных маркеров для построения диагностической модели на основе логистической регрессии.

Степень достоверности: Достоверность полученных результатов обеспечена использованием в исследовании образцов наборов данных для нескольких различных реальных клинических исследований и корректностью применения апробированного в научной практике исследовательского и аналитического аппарата.

Апробация диссертации: Основные результаты работы были представлены на международных и российских конференциях: The 5th Annual European Congress of The Association for Mass Spectrometry: Applications to the Clinical Lab (MSACL

2018 EU 5rd Annual Congress), Salzburg, Austria September 11-13, 2018; 4-й междисциплинарный научный форум с «Новые материалы и перспективные технологии», Москва, Россия, 27-30 ноября 2018 года; 9-й съезд ВМСО VIII Всероссийской конференции с международным участием «Масс-спектрометрия и ее прикладные проблемы», Россия, Москва, 14–18 октября 2019 года.

Личный вклад автора заключался в изучении и систематизации литературы по данной тематике, выборе методов обработки масс-спектрометрических данных и проведению обработки, а также выбора критериев для анализа и интерпретации полученных результатов. Было принято участие в подготовке научных статей, опубликованных в соавторстве.

Структура и объём диссертации. Диссертация состоит из введения, 4 глав, заключения, выводов, списка литературы. Работа изложена на 118 страницах машинописного текста, содержит 17 таблиц, 15 рисунков. Библиографический указатель содержит 171 источник, все относятся к зарубежным.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность работы, делается оценка разработанности темы, сформулированы цели и задачи, продемонстрирована новизна и практическая значимость.

Первая глава (Обзор литературы) содержит разделы, в которых: рассматривается структура липидов, характерные особенности классов липидов и их роль в метаболизме человека; методы масс-спектрометрического анализа липидов, в число которых вошли масс-спектрометрия с прямой ионизацией, масс-спектрометрия прямого ввода, масс-спектрометрия с предварительным хроматографическим разделением, их достоинства, недостатки и случаи применения в клинических исследованиях; методы предобработки и обработки данных, полученных в ходе масс-спектрометрического анализа, включающие в

себя фильтрацию шумов, выделение пиков, деизотопирование, выравнивание, заполнение пропусков, нормализация данных, идентификация липидов, построение классифицирующей модели и выбор соединений-биомаркеров и частоту их применения при исследованиях характера «случай/контроль».

Вторая глава содержит разделы, в которых указаны материалы, задействованные в исследованиях, и описаны методы исследования.

В разделе 2.1 перечислены биологические образцы, задействованные в исследованиях и собранные в ФГБУ «Национальный исследовательский центр акушерства, гинекологии и перинатологии им. В. И. Кулакова»:

- ткани кисты яичника, полученные в ходе лапароскопической операции (набор 1);
- ткани молочной железы и опухоли молочной железы, взятые у 25 человек (набор 2);
- ткани опухоли шейки матки, взятые у 33 человек и окружающей здоровой ткани, взятые у 25 человек (набор 3);
- ткани молочной железы и опухоли молочной железы, взятые у 40 человек, не имеющих метастазы в регионарные лимфоузлы и у 48 человек, имеющих местастазы в регионарные лимфоузлы (набор 4);
- ткани эндометрия, взятые у 30 человек с миомой в секреторную фазу, у 30 человек с эндометриозом в секреторную фазу, у 30 человек с миомой в пролиферативную фазу, у 30 человек с эндометриозом в пролиферативную фазу (набор 5);
- образцы плазмы крови от 40 здоровых людей, от 36 людей с диагностированной внутриутробной задержкой развития плода, от 28 людей с диагностированной преэклампсией (набор 6);
- образцы плазмы крови от 32 здоровых людей, 46 людей с впервые диагностированной миомой, 60 людей с рецидивом миомы (набор 7);

- образцы плазмы крови от 36 людей с эндометриозом и 24 здоровых людей (набор 8).

Кроме того, в разделе приведено описание химических препаратов, задействованных в пробоподготовке образцов, проведении масс- и хромато-масс-спектрометрического анализа, а также синтезе стандартов липидов PC 18:1/18:1 и PC 18:1/20:4.

В разделе 2.2 описана подготовка образцов для масс-спектрометрического анализа: приведён алгоритм для синтеза стандартов липидов PC 18:1/18:1 и PC 18:1/20:4 и приготовления смесей стандартов с различным относительным содержанием липидов; описан алгоритм экстракции липидов из образцов наборов 3-8 на основе метода Фолча; описано разбиение образцов из наборов 4-7 на партии для хромато-масс-спектрометрического анализа:

Набор данных 4 был разбит на 3 партии:

- 16 образцов из группы 1, 14 образцов из группы 2, 16 образцов из группы 3, 14 образцов из группы 4;
- 10 образцов из группы 1, 20 образцов из группы 2, 10 образцов из группы 3, 20 образцов из группы 4;
- 14 образцов из группы 1, 14 образцов из группы 2, 14 образцов из группы 3, 14 образцов из группы 4,

где к группе 1 относились ткани опухоли молочной железы от пациентов без метастаз, к группе 2 – ткани опухоли молочной железы от пациентов с метастазами, к группе 3 – ткани молочной железы от пациентов без метастаз, к группе 4 – ткани молочной железы от пациентов с метастазами.

Набор данных 5 был разбит на 5 партий: каждая партия содержала по 6 образцов ткани из каждой клинической группы.

Набор данных 6 был разбит на 4 партий:

- 10 образцов из группы 1, 10 образцов из группы 2, 10 образцов из группы 3;

- 10 образцов из группы 1, 10 образцов из группы 2, 10 образцов из группы 3;
- 10 образцов из группы 1, 12 образцов из группы 2, 8 образцов из группы 3;
- 10 образцов из группы 1, 4 образца из группы 2,

где к группе 1 относились образцы плазмы от пациентов контрольной группы, к группе 2 – образцы плазмы от пациентов с диагнозом «задержка развития плода», к группе 3 – образцы плазмы от пациентов с преэклампсией.

Набор данных 7 был разбит на 8 партий:

- 6 образцов из группы 1;
- 2 образца из группы 1, 10 образцов из группы 2, 10 образцов из группы 3;
- 4 образца из группы 1, 4 образца из группы 2, 10 образцов из группы 3;
- 4 образца из группы 1, 6 образцов из группы 2, 10 образцов из группы 3;
- 4 образца из группы 1, 6 образцов из группы 2, 10 образцов из группы 3;
- 2 образца из группы 1, 8 образцов из группы 2, 10 образцов из группы 3;
- 6 образцов из группы 1, 4 образца из группы 2, 10 образцов из группы 3;
- 4 образца из группы 1, 2 образца из группы 2,

где к группе 1 относятся образцы плазмы крови от пациентов контрольной группы, к группе 2 – образцы плазмы крови от пациентов с впервые диагностированной миомой, к группе 3 – образцы плазмы крови от пациентов с рецидивом миомы.

Набор данных 8 был разбит на две партии, каждая содержала по 18 образцов из группы 1 и 12 образцов из группы 2.

В разделе 2.3 описаны методы масс-спектрометрического анализа, задействованные при выполнении работы: масс-спектрометрия с прямой ионизацией на приборах 7T LTQ FT Ultra для набора 1 и Q-TOF Maxis Impact для набора 1 и 2; масс-спектрометрия прямого ввода на приборах 7T LTQ FT Ultra для

смесей синтезированных липидов и Q-TOF Maxis Impact для смесей синтезированных липидов и экстрактов образцов набора 3; масс-спектрометрия с предварительным разделением на обратно-фазовой хроматографической колонке Zorbax C18 и с использованием хроматографа Dionex UltiMate 3000 и масс-спектрометра Q-TOF Maxis Impact для экстрактов образцов из наборов 4-8.

<u>В разделе 2.4</u> описаны методы предобработки и обработки данных, полученных в ходе масс-спектрометрического анализа. Указано, что для редуцирования отклонений между партиями наборов 4-7 использовались следующие методы нормализации данных:

- автомасштабирование: $I^*_{p,s,b} = \frac{I_{p,s,b} - <I_{p,b}>}{sd(I_{p,b})} * sd(I_p) + < I_p >$;
- паретомасштабирование: $I^*_{p,s,b} = \frac{I_{p,s,b} - <I_{p,b}>}{\sqrt{sd(I_{p,b})}} * \sqrt{sd(I_p)} + < I_p >$;
- масштабирование на диапазон: $I^*_{p,s,b} = \frac{I_{p,s,b} - <I_{p,b}>}{\max(I_{p,b}) - \min(I_{p,b})} * (\max(I_p) - \min(I_p)) + < I_p >$;
- масштабирование на уровень: $I^*_{p,s,b} = \frac{I_{p,s,b} - <I_{p,b}>}{<I_{p,b}>} * < I_p > + < I_p >$;
- квантильная нормализация. Каждая клиническая группа подвергалась нормализации отдельно;
- VSN – каждая клиническая группа подвергалась нормализации отдельно;
- PQN – в исследовании использовались два типа референсных спектров. Первый тип референса рассчитывался как среднее значение интенсивности пиков по всем образцам и обозначался как «PQN с исс-ых спектров реф.». Второй рассчитывался как средние значения пиков по спектрам контроля качества и обозначался как «PQN с КК реф.»,

где p – номер пика, s – номер образца, b – номер партии, $<I_{p,b}>$ и $sd(I_{p,b})$ – среднее значение и стандартное отклонение площади пика p по партии b, $<I_p>$ и $sd(I_p)$ – среднее значение и стандартное отклонение площади пика p по всем образцам.

$I_{p,s,b}$ и $I^*_{p,s,b}$ – значения площади пика p из образца s партии b до и после нормализации.

При сравнении эффективности различных наборов переменных для построения классификационных моделей для набора данных 2 были созданы следующие наборы данных: все зарегистрированные соединения; соединения со статистически значимой разницей в интенсивности между спектрами от здоровой и от опухолевой ткани; соединения, идентифицированные как липиды; соединения со статистически значимой разницей в интенсивности между спектрами от здоровой и от опухолевой ткани, идентифицированные как липиды.

Аналогичные наборы переменных были созданы на основе 10 образцов из набора 3 (набор 3a). Кроме того, были созданы моноклассовые наборы переменных – наборы, куда входили липиды только из одного класса. Для указанных наборов данных были построены классификационные OPLS-DA модели.

Для OPLS-DA классификации образцов относительно диагностических задач, сформулированных на основе наборов данных 3, 4, 7, 8, были созданы наборы переменных исходя из наличия статистической разницы в уровнях между «нормальным» образцом и «патологическим», значения проекций переменных в OPLS-DA модели (со значением переменной больше 1), используя $\sqrt{N}$ переменных с максимальным Gini индексов в случайном лесе из N всех переменных, со значениями, превышающими 0,05 веса переменной с максимальным весом в машине опорных векторов и моноклассовых наборов переменных (рисунок 1a, 2).

Из выбранных наборов переменных и исходных наборов переменных для построения диагностической модели на основе логистической регрессии были отобраны переменные согласно информационному критерию Акаике: переменные поэтапно добавлялись в логистическую регрессию, пока наблюдался рост критерия Акаике. После чего поэтапно исключались переменные, у чьих

коэффициентов вероятность равенства нулю была больше 0,05. Также для выбора маркеров из исходных наборов переменных и сформулированных с использованием теста Манна-Уитни был использовано ЛАССО, с последующим исключением переменных (рисунок 1а, 2).

Проверка качества моделей осуществлялась с использованием тестовых образцов (набор 2), кросс-валидации по отдельному объекту и вложенной кросс-валидации под отдельному объекту (набор 3,4, 7, 8) (рисунок 1, 2).

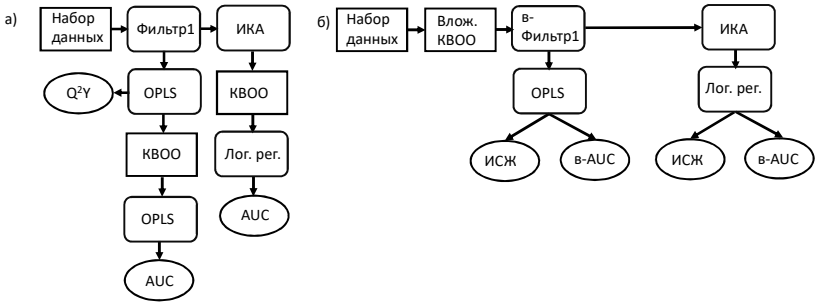

Рисунок 1. а) Блок-схема обработки данных для сравнения стратегий выбора переменных для моделей на основе OPLS и логистических регрессии. б) Блок-схема обработки данных для сравнения стабильности методов селекции переменных. Фильтр1 – фильтр первого уровня: фильтрация не осуществлялась, Манна-Уитни, МОВ, СЛ, OPLS или по выбранным классам липидов; в-Фильтр1 – фильтр первого уровня: фильтрация не осуществлялась, Манна-Уитни, МОВ, СЛ или OPLS; ИКА-фильтр второго уровня, OPLS – модель на основе дискриминантного анализа OPLS; Лог. рег – модель на основе логистической регрессии; влож. КВОО – разбиение данных для вложенной кросс-валидации по отдельному объекту; КВОО – разбиение данных для кросс-валидации по отдельному объекту; Q^2Y – параметр предсказательной способности OPLS модели; AUC – площадь под операционной кривой, полученная в ходе кросс-валидации; в-AUC – площадь под операционной кривой, полученная в ходе вложенной кросс-валидации; ИСЖ – индекс сходства Жаккара.

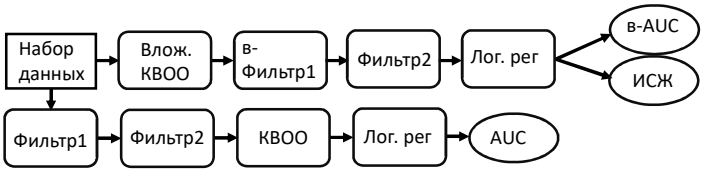

Рисунок 2. Блок-схема обработки данных для сравнения ЛАССО и ИКА фильтров для создания модели на основе логистической регрессии. Фильтр1 – фильтр первого уровня: отсутствие фильтра или фильтр Манна-Уитни; в-Фильтр1 – фильтр первого уровня, применённый в ходе вложенной кросс-валидации; Фильтр2 – фильтр второго уровня: ИКА или ЛАССО; Лог. рег – модель на основе логистической регрессии; Влож. КВОО – разбиение данных для вложенной кросс-валидации по отдельному объекту; КВОО – разбиение данных для кросс-валидации по отдельному объекту; AUC – площадь под операционной кривой, полученная в ходе кросс-валидации; в-AUC – площадь под операционной кривой, полученная в ходе вложенной кросс-валидации; ИСЖ – индекс сходства Жаккара.

Для оценки эффективности методов нормализации рассчитывалось 4 параметра: дистанция между кластерами партий (ДП) и дистанция между кластерами клинических групп (ДГ) в пространстве главных компонент; число соединений со статистически значимой разницей площадей пиков между партиями (СП) и число соединений со статистически значимой разницей площадей пиков между клиническими группами (СГ) (рисунок 3).

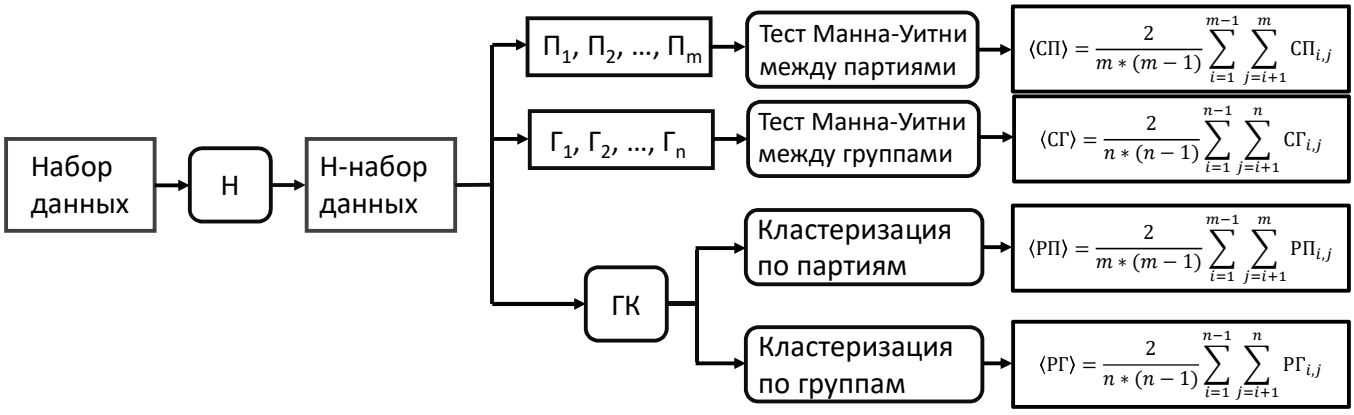

Рисунок 3. Блок-схема обработки данных при сравнении методов нормализации. Н – один из использованных методов нормализации (нет нормализации,

автошкалирование, Парето шкалирование, шкалирование на диапазон, шкалирование на уровень, квантильная нормализация, PQN, VSN); Н-Набор Данных – набор данных после нормализации методом Н; $П_1$, $П_2$, …, $П_m$ – партии анализа в наборе данных; m – число партий в наборе данных; $Г_1$, $Г_2$, …, $Г_n$ – клинические группы в наборе данных; n –число клинических групп в наборе данных; ГК – трансформация набора данных в пространство главных компонент; $РП_{ij}$, $РГ_{ij}$ – расстояние между партиями-кластерами и расстояние между группами-кластерами i и j; $СП_{ij}$, $СГ_{ij}$ – число соединений со статистически значимой разницей в уровнях между партией/группой i и j; <СП>, <СГ>, <РГ>, <РП> – среднее соответствующей метрики в заданной комбинации набора данных и метода нормализации.

Третья глава содержит результаты, полученные при исследовании масс-спектров фосфолипидов, полученных с использованием масс-спектрометрии с прямой ионизацией, сравнительного анализа методов нормализации данных и методов выбора переменных для построения классификационной и диагностической моделей.

В разделе 3.1 на примере липидов PC 18:1/18:1(PC1) и PC 18:1/20:4 (PC2) показана возможность деконволюции перекрывающихся сигналов от [PC x:y + $Na^+]^+$ и [PC (x+2):(y+3) + $H^+]^+$ на основе интенсивностей характерных спектров фрагментации: натрий-иона фосфохолина m/z 147,0 и иона фосфотидилхолина m/z 184,1. Предположение, что интенсивность пиков характерных ионов-фрагментов в тандемном спектре пропорциональна количеству соответствующих ионов-прекурсоров, делает возможным формирование следующих уравнений:

$$I_{[PC1+N\ ^+]^+} = \frac{I_2}{I^F_{[PC1+Na^+]^+} + I^F_{[PC2+H^+]^+}} I^F_{[PC1+Na^+]^+} \quad (1)$$

$$I_{[PC2+H^+]^+} = \frac{I_2}{I^F_{[PC1+Na^+]^+} + I^F_{[PC2+H^+]^+}} I^F_{[PC2+H^+]^+} \quad (2),$$

где $I_{[PC1+Na^+]^+}$ и $I_{[PC2+H^+]^+}$ - интенсивность сигналов от перекрывающихся натрий-иона и протонированного иона, соответственно, I_2 – результирующий

сигнал $I_{[PC1+\ \ ^+]^+} + I_{[PC2+H^+]^+}$, $I^F_{[PC1+N\ ^+]^+}$ и $I^F_{[PC2+\ ^+]^+}$, – интенсивность характерных ионов m/z 147.0 и m/z 184.1 соответственно. То, что интенсивность пиков в режиме полного сканирования пропорциональна концентрации соответствующих соединений в образце и ионизация фосфотидилхолинов при масс-спектрометрии прямой ионизации происходит присоединением протона или натрия, позволяет написать

$$\frac{I_{PC2}}{I_{PC1}} = \frac{I_{[PC2+H^+]^+} + I_{[PC2+Na^+]^+}}{I_{[PC1+H^+]^+} + I_{[PC1+N\ ^+]^+}} \quad (3),$$

где $I_{[PC1+H^+]^+}$ и $I_{[PC2+N\ ^+]^+}$ - интенсивности сигнала от протонированного PC1 иона и натрий-иона PC2.

Проверка релевантности данных уравнений по результатам масс-спектрометрического анализа созданных растворов PC1 и PC2 на LTQ-FT Ultra в режиме полного сканирования и по спектрам фрагментации, полученным с использованием Maxis Impact показала хорошее согласие данных, полученных на LTQ-FT Ultra и Maxis Impact между собой и с реальным отношением химического количества, а также корреляцию с последним с коэффициентом 0.997 (доверительный интервал 0,99 – 1,00).

Анализ масс-спектров, полученных при масс-спектрометрическом анализе методом прямой ионизации тканей показал высокий уровень корреляции значений отношений концентраций соединений, вычисленных каждым из рассмотренных методов – 0,996 (ДИ 0,99 – 1,00).

В разделе 3.2 приведены результаты сравнительного анализа методов нормализации данных на основе исследуемых данных. Для всех ненормализованных наборов данных, значения РП были либо статистически значимо больше, чем значения РГ, либо статистически значимой разницы в их значениях не было. Применение паретомасштабирования и автомасштабирования привело к тому, что РГ стали статистически значимо больше, чем РП в 7 случаях из 8. VSN и PQN не показали статистически значимого увеличения РГ

относительно ДП по сравнению с ненормализованными данными. РП значительно сократилось после автомасштабирования, паретомасштабирования и масштабирования на уровень в 7 из 8 случаях, в то время как ДГ после автомасштабирования и паретомасштабирования не изменилось статистически значимо во всех случаях.

В большинстве исследуемых случаев для ненормализованных данных число СП статистически значимо больше, чем число СГ. Автомасштабирование приводит к возрастанию значения СГ для всех анализируемых случаев. PQN и VSN не увеличивают число наборов данных, для которых СГ статистически значимо больше СП.

Среднее число СП в ненормализованных данных, после квантильной нормализации, PQN и VSN не имеют между собой статистически значимых отличий. Число СП в данных после автомасштабирования, паретомасштабирования, масштабирования на уровень и масштабирования на диапазон статистически значимо меньше, чем в ненормализованных данных ($p = 0,007$).

Число СГ в данных после автомасштабирования и квантильной нормализации статистически значимо выше, чем в ненормализованных данных ($p = 0,04$ и $p = 0,007$ для каждой пары соответственно), в то время как после PQN число СГ статистически значимо ниже, чем в ненормализованных данных ($p = 0,007$).

В разделе 3.3 приведены результаты сравнительного анализа методов выбора переменных для OPLS-DA моделей и для моделей на основе логистической регрессии.

При анализе набора 2 было выявлено, что большая часть липидов, имеющая статистически значимую разницу в уровнях между нормальной и опухолевой тканью, относилась к классам сфингомиелинов, фосфатидилхолинов и фосфатидилинозитолов. Наибольшие значения Q^2Y встречаются в моделях,

построенных с использованием соединений со статистически значимой разницей в уровнях и с использованием всех соединений. При построении модели на основе липидов наблюдается рост параметра R^2X по сравнению с моделями, где были задействованы все соединения. Следует отметить значительное ухудшение качества моделей в режиме отрицательных ионов относительно режима положительных ионов.

Для набора 3а в режиме положительных ионов наилучший ожидаемый предсказательный потенциал продемонстрировала модель, построенная на наборе неполярных глицеролипидов (Q^2Y = 0,64, AUC = 0,95). Модель, основанная на фосфотидилэтаноламинах, показала хороший предсказательный потенциал (Q^2Y = 0,48, AUC =0,86).

Модели, построенные на основе одного класса липидов, удовлетворяли условию $Q^2Y > 0,4$ в большем числе случаев, чем остальные методы (4 из 14): диагностика рецидива миомы в режиме отрицательных ионов, диагностика рака шейки матки в режиме положительных и в режиме отрицательных ионов, диагностика эндометриоза в режиме положительных ионов. Данные модели были построены на основе фосфатидилэтаноламинов (PE), фосфатидилхолинов (PC), фосфатидилсеринов (PS), сфингомиелинов (SM) и липидов с простой эфирной связью.

Для всех наборов данных, сформированных при помощи описанных фильтров, кроме тех, кто проводил выборку на основе значений ПП и с использованием МОВ с сигмоидальным и полиномиальным ядром, была зафиксирована статистически значимая разница между AUC, полученной в ходе кросс-валидации и AUC, полученной в ходе вложенной кросс-валидации. Наибольшая разница наблюдается у данных, полученных с использованием теста Манна-Уитни, и СЛ.

OPLS модели, которые строились на основе отдельных классов липидов, показали наилучшее значение AUC относительно других методов в 12 случаях из

14, и в 6 случаях значение AUC превышало 0.8. Тест Манна-Уитни показал статистически значимо меньшую ($p < 0{,}001$) AUC для OPLS моделей, чем для тех, для построения которых использовалась логистическая регрессия, в случае проведения исследования по схеме, представленной на рисунке 2а. При использовании вложенной кросс-валидации (рисунок 2б) различия не являлись статистически значимыми ($p = 0{,}18$).

Комбинация СЛ с ИКА даёт статистически значимо меньшие значения ИСЖ, чем остальные методы. B-AUC статистически значимо меньше AUC для всех методов, использованных для выбора переменных для логистической регрессии.

Число случаев в ходе вложенной кросс-валидации, когда построение логистической модели было невозможно, было больше при том способе фильтрации, когда использовалось ЛАССО, по сравнению с ИКА. Регрессионная модель не была создана в ходе вложенной кросс-валидации в ходе как минимум одной итерации для 9 задач в случае ЛАССО селекции и для 4 задач в случае комбинации ЛАССО селекции с тестом Манна-Уитни. В то же время, создание регрессионной модели не было возможно в ходе как минимум одной итерации вложенной кросс-валидации для 4 задач при чистой ИКА селекции и в случае комбинации ИКА селекции с тестом Манна-Уитни – для 2 задач. Только фильтр по ПП в качестве первого фильтра для всех задач на всех итерациях вложенной кросс-валидации даёт логистическую регрессию со статистически значимыми ненулевыми коэффициентами.

Четвёртая глава содержит обсуждение результатов исследования с учётом работ, проведённых другими исследователями в данной области. Исследователи, решающие проблему деконволюции перекрывающихся сигналов от липидов при масс-спектрометрическом анализе без хроматографического разделения, выделяют следующие случаи перекрытия: $[\mathrm{L\,x:y + Ad}]^{Z}$ ^{2} ^{13}C и $[\mathrm{L\ x{:}(y\text{-}1) + Ad}]^{Z}$ [154], $[\mathrm{L_1\ x_1{:}y_1 + Ad_1}]^{Z}$ и $[\mathrm{L_2\ x_2{:}y_2 + Ad_2}]^{Z}$, $[\mathrm{L\ x^1{:}y^1\_\ x^2{:}y^2 + Ad}]^{Z}$ и $[\mathrm{L\ x^3{:}y^3\_\ x^4{:}y^4 +}$

$Ad]^Z$ [155], где L, L_1 и L_2 – обозначение класса липида; x, x_1 и x_2 – общее число атомов углерода в жирнокислотных остатках; y, y_1 и y_2 – общее число двойных связей в жирнокислотных остатках; x^1, x^2, x^3, x^4 – число атомов углерода в жирнокислотных остатках, y^1, y^2, y^3, y^4 – число двойных связей в жирнокислотных остатках; Ad, Ad_1 и Ad_2 – аддукт, несущий заряд; Z – заряд; $2^{13}C$ – два атома углерода в атоме являются изотопами ^{13}C. В литературе для деконволюции второго случая (частным случаем которого является PC x:y + $Na^+]^+$ и [PC (x + 2) : (y + 3) + $H^+]^+$) предлагаются варианты использования внутренних стандартов и дериватизации липидов в ходе пробоподготовки. Использование данных методов представляет сложности в случае масс-спектрометрии с прямой ионизацией. Созданный метод потенциально применим в третьем случае при анализа липидов в режиме отрицательных ионов, где ионы-фрагменты жирнокислотных остатков являются характерными ионами для родительского иона и относительная доля каждого липида будет вычисляться исходя из интенсивностей ионов жирнокислотных остатков. Теоретически возможна деконволюция первого примера, на основе значений интенсивности характерных фрагментов, не содержащих ^{13}C и содержащих атом ^{13}C, но практически затруднена из-за низкой интенсивности фрагмента, содержащего ^{13}C от $[L\ x : y + Ad]^{Z^{2^{13}C}}$.

Предполагается, что низкая эффективность PQN связана с тем, что метод использует медианное значение всех пиков в спектре или хроматограмме и не должен уменьшать межгрупповые различия, которые вызваны небольшим количеством соединений в образце. Поэтому, если эффекты между партиями вызваны относительно небольшим количеством пиков или имеют немонотонный характер (некоторые интенсивности увеличиваются, некоторые уменьшаются от образца к образцу), PQN может быть неэффективным.

VSN основан на модели, представляющей сигнал в виде $X = \alpha + \mu e^{\eta} + \varepsilon$, где X – измеренное значение интенсивности компонента, α – смещение по интенсивности, μ – значение интенсивности компонента с удалённым шумом, ε и

η – аддиктивная и мультипликативная ошибки. Это даёт возможность представить среднее значение и дисперсию X как $E\,(X) = \alpha + m_\eta \mu$ и $Var(X) = s_\eta^2 \mu^2 + \sigma_\varepsilon^2$, где m_η и s_η^2 – среднее значение и дисперсия e^η соответственно, и σ_ε – стандартное отклонение ε, которое рассчитывается в ходе анализа спектров образцов на основе сигнала фона. Этот метод успешно используется в ПЦР и ЯМР анализе, но при ЖХ-МС анализе стандартное отклонение ошибки может быть определено неправильно, поскольку при некоторых методах предобработки для ускорения анализа исключаются пики, которые могут быть использованы в качестве шума. Это могло стать причиной низкой эффективности VSN в данном исследовании.

Для классификационных OPLS-DA моделей, построенных без не-липидов или соединений, не удовлетворяющих наличию статистически значимой разницей в уровнях по тесту Манна-Уитни, для построения классификационной модели привело к росту параметра R^2X, характеризующего степень описания независимых переменных построенной моделью. Это позволяет говорить о том, что ограничение анализа липидами приводило к уменьшению ошибок при поиске соединений, ключевых для построения дискриминантной модели.

Выбор одного класса липидов повысил качество классификационной модели в отдельно взятых случаях: фосфатидилэтаноламины были эффективны в дискриминации рецидива миомы и контрольной группы по плазме крови; они же, липиды с простой эфирной связью и фосфатидилсерины эффективно разделяли раковую ткань шейки матки и здоровую ткань; сфингомиелины эффективно дифференцировали плазму крови от женщин с эндометриозом и без. Кроме того, липиды, которые внесли наибольший вклад в разделение опухолевой и здоровой ткани молочной железы, преимущественно относились к классам сфингомиелинов, фосфатидилхолинов и фосфатидилинозитолов.

Вышеупомянутые группы липидов являются ключевыми для ряда процессов клеточного метаболизма: уровень фосфатидилэтаноламинов и фосфоэтаноламинов, стимулирующих рост клеток, в злокачественных тканях по сравнению с нормальными повышается. Увеличение количества

фосфатидилсерина во внешних мембранах клеток является сигналом для начала апоптотических процессов. Липиды с простой эфирной связью являются антиоксидантами и задействованы в транспорте через мембрану, сфингомиелины задействованы в противовоспалительных процессах и транспорте холестерина.

К достоинствам классификационных моделей на основе отдельных классов относится независимость от набора образцов и создаваемый таким образом набор переменных обладает наибольшей, по сравнению с остальными методами, устойчивостью.

Получено, что метод выбора переменных с использованием информационного критерия Акаике показал большую устойчивость, чем использование ЛАССО. Обсуждая этот факт, следует сделать акцент на том, что ИКА ставит более жёсткое ограничение на число переменных N ($2\ln\big(L(\beta)\big) - 2N$), чем ЛАССО, в то время как ЛАССО $\ln\big(L(\beta)\big) - \lambda\sum_{i=1}^{N}|\beta_i|$ позволяет большее количество переменных в модели за счёт уменьшения абсолютного значения коэффициентов $|\beta_i|$ при переменных. Уменьшение абсолютного значения коэффициентов затрудняет признание коэффициентов статистически значимыми по ненулевой гипотезе и обуславливает меньшее количество случаев, когда модель была создана по сравнению с методами, когда был задействован ИКА. В работе Lee «Robustness of chemometrics-based feature selection methods in early cancer detection and biomarker discovery» селекция по проекции переменной была показана более устойчивой, чем селекция по t-тесту.

ЗАКЛЮЧЕНИЕ

В ходе исследования получен метод, позволяющий проводить деконволюцию перекрывающихся изобарических пиков на основе информации, полученной при тандемной масс-спектрометрии без использования дополнительных реагентов или введения внутренних стандартов. Данный метод основан на вычислении отношений интенсивности характерных ионов фрагментации от перекрывающихся родительских ионов. Также проведён сравнительный анализ методов нормализации на основе данных исследуемых образцов и показано, что автомасштабирование сохраняет статистически значимые различия между диагностическими группами, нивелируя различия между партиями. Для более широко используемых PQN и VSN это подтверждено не было, что связывается с характером колебаний значений площадей пиков в случае PQN и методами предобработки хромато-масс-спектрометрических данных в случае VSN. В предложенном варианте автомасштабирования задействуются средние значения и стандартные отклонения значений площадей пиков по отдельным партиям и по всему анализируемому набору. Проведена сравнительная оценка эффективности построения дискриминационных моделей по полному спектру, идентифицированным липидам и отдельным классам липидов и показана потенциальная эффективность построения дискриминационных и диагностических моделей по отдельным классам липидов. Исходя из качества построенных моделей на основе отдельных классов липидов можно делать предположения о патофизиологии заболевания. Сравнительная оценка площадей под ROC кривой при кросс-валидации по отдельному объекту и вложенной кросс-валидации по отдельному объекту показала большую устойчивость OPLS-DA моделей по сравнению с моделями на основе логистической регрессии, что позволяет использовать OPLS-DA для больших наборов переменных (например, упомянутые выше классы липидов). При оценке стабильности метода выбора переменной с использованием вложенной кросс-валидации, было получено, что двухстадийный фильтр по значениям проекций переменных в OPLS-DA модели и информационному критерию Акаике в

логистической регрессии показывает наилучшие результаты. Предлагается выполнять поиск переменных в два этапа: вычисление значений проекций переменной методами, используемыми для построения OPLS-DA модели на основе всех зарегистрированных и /или идентифицированных соединений; поэтапный выбор переменных из отобранных по значению проекции переменной в логистическую регрессию, пока наблюдается увеличение информационного критерия Акаике для модели на основе данной логистической регрессии.

ВЫВОДЫ

1. Деконволюция перекрывающихся протонированных и натрий-ионизированных фосфолипидов при масс-спектрометрии с прямой ионизацией на основе сигнала от протонированного более лёгкого липида (PC1) ($I_{[PC1+H]^+}$), натрий-иона более тяжёлого липида (PC2) $I_{[PC2+Na]^+}$, сигнала от перекрывающихся натрий-иона PC1 и протонированного липида PC2 (I_2), а также характерного иона от протонированного иона PC2 184,1 m/z ($I^F_{[PC2+H]^+}$) и от натрий-иона PC1 147,0 m/z ($I^F_{[PC1+Na]^+}$) и формул $I_{[PC1\quad]^+} = \frac{I_2}{I^F_{[PC1+Na]^+}+I^F_{[PC2+H]^+}} I^F_{[PC1+Na]^+}$, $I_{[PC2+H]^+} = \frac{I_2}{I^F_{[PC1+Na]^+}+I^F_{[PC2+H]^+}} I^F_{[PC2+H]^+}$, $\frac{I_{PC2}}{I_{PC1}} = \frac{I_{[PC2+H]^+}+I_{[PC2+Na]^+}}{I_{[PC1+H]^+}+I_{[PC1+\quad]^+}}$ позволяет с высокой точностью определить соотношение липидов в образце.

2. Автомасштабирование продемонстрировало более высокую эффективность при обработке данных для клинических исследований, чем более широко распространенные в молекулярном профилировании методы вероятностной нормализации на частное и нормализации стабилизацией дисперсии. Предлагается использовать метод нормализации на основе автомасштабирования $I^*_{p,s,b} = \frac{I_{p,s,b}-<I_{p,b}>}{sd(I_{p,b})} * sd(I_p)+< I_p >$, где p – номер пика, s – номер образца, b – номер партии, $<I_{p,b}>$ и $sd(I_{p,b})$ – среднее значение и стандартное отклонение площади пика p по партии b, $<I_p>$ и $sd(I_p)$ – среднее значение и стандартное

отклонение площади пика p по всем образцам. $I_{p,s,b}$ и $I^*_{p,s,b}$ – значения площади пика p из образца s партии b до и после нормализации.

3. Широко используемый при поиске биомаркеров метод ЛАССО показал более низкую устойчивость, чем поиск по информационному критерию Акаике.

4. Предлагается двухстадийный метод поиска биомаркеров для логистической регрессии при омиксовых исследованиях: с использованием ортогональных проекций на скрытые структуры вычисляются значения проекции переменной для соединений, задействованных в классификации и выбираются соединения с ПП >1. Среди них поэтапно отбираются в модель на основе лог. регресии переменные, пока растёт критерий Акаике.

РАБОТЫ, ОПУБЛИКОВАННЫЕ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Chagovets, V., Kononikhin, A., **Tokareva, A.**, Bormotov, D., Starodubtseva, N., Kostyukevich, Y., Popov, I., Frankevich. V., Nikolaev, E. Relative quantitation of phosphatidylcholines with interfered masses of protonated and sodiated molecules by tandem and Fourier-transform ion cyclotron resonance mass spectrometry// European Journal of Mass Spectrometry. 2019. Vol. 25 № 2. P. 259-264.
2. **Tokareva, A. O.**, Chagovets V. V., Kononikhin, A. S., Starodubtseva, N. L., Nikolaev, E. N., Frankevich, V. E. Normalization methods for reducing interbatch effect without quality control samples in liquid chromatography-mass spectrometry-based studies // Analytical and Bioanalytical Chemistry. 2021. Vol. 413 № 13. P. 3479-3486.
3. **Tokareva, A. O.**, Chagovets V. V., Starodubtseva, N. L., Nazarova, N. M., Nekrasova, M. E., Kononikhin, A. S., Frankevich. Feature selection for OPLS discriminant analysis of cancer tissue lipidomics data // Journal of Mass Spectrometry. 2020. Vol. 55 № 1. № e4457

4. Chagovets, V. V., Starodubtseva, N. L., **Tokareva A. O.**, Frankevich V. E., Rodionov V. V., Kometova V. V., Chingin, K., Kukaev, E. N., Chen, H., Sukhikh, G. T. Validation of breast cancer margins by tissue spray mass spectrometry // International Journal of Molecular Science. 2020. Vol. 21 № 12. № 4568
5. **Tokareva, A. O.**, Chagovets, V. V., Kononikhin, A. S., Starodubtseva, N. L., Nikolaev, E. N., Frankevich, V. E. Comparison of the effectiveness of variable selection method for creating a diagnostic panel of biomarkers for mass spectrometric lipidome analysis // Journal of Mass Spectrometry. 2021. Vol. 56 № 3. № e4702
6. **Токарева А. О.,** Чаговец В. В., Кононихин А. С., Стародубцева Н. Л., Франкевич. В. Е., Николаев Е. Н. Алгоритм обработки масс-спектрометрических данных для получения диагностической панели молекулярных соединений на примере поиска маркеров метастазирования при раке молочной железы // Biomedical Chemistry: Research and Methods. 2021. Vol 4 № 3. № e00156.
7. **Токарева А. О.**, Чаговец В. В., Кононихин А. С., Стародубцева Н. Л., Франкевич В. Е., Николаев Е. Н. Способ обработки масс-спектрометрических данных для повышения эффективности поиска диагностических маркеров при проведении клинических исследований. Заявка на патент РФ № 2021126242 от 07.09.2021