МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.ЛОМОНОСОВА

На правах рукописи

ТУРОВА ПОЛИНА НИКОЛАЕВНА

НОВЫЕ СПОСОБЫ ОБРАБОТКИ ХРОМАТОМАСС-СПЕКТРОМЕТРИЧЕСКИХ ДАННЫХ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОИСКА БИОМАРКЕРОВ И КЛАССИФИКАЦИИ РАСТЕНИЙ

Специальность – 02.00.02 — Аналитическая химия

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата химических наук

Москва — 2022

Работа выполнена на кафедре аналитической химии Химического факультета МГУ имени М.В. Ломоносова.

Научный руководитель **Ставрианиди Андрей Николаевич**

Кандидат химических наук

Официальные оппоненты: **Кирсанов Дмитрий Олегович**

Доктор химических наук

Институт химии ФГБОУ ВО «Санкт-Петербургский государственный университет», профессор

Темердашев Азамат Зауалевич

Доктор химических наук

ФГБОУ ВО «Кубанский Государственный Университет», старший научный сотрудник

Борисов Роман Сергеевич

Кандидат химических наук

ФГБУН ордена Трудового Красного Знамени «Институт нефтехимического синтеза им. А. В. Топчиева РАН», ведущий научный сотрудник

Защита диссертации состоится «27» апреля в 15 часов 00 минут на заседании диссертационного совета МГУ.02.05 Московского государственного университета имени М.В.Ломоносова по адресу: 119991, г. Москва, ГСП-1, Ленинские горы, д.1, стр. 3, МГУ имени М.В.Ломоносова, Химический факультет, аудитория 337.

E-mail: dissovet02.00.02@mail.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В.Ломоносова (Ломоносовский просп., д. 27) и на сайте ИАС «ИСТИНА»: https://istina.msu.ru/dissertations/438983253/

Автореферат разослан «22» марта 2022 года.

Ученый секретарь
диссертационного совета,
кандидат химических наук И.А. Ананьева

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Одной из актуальных задач, решаемых с помощью хроматомасс-спектрометрических методов анализа является классификация и контроль качества растительных материалов и продуктов на их основе. Такие продукты могут содержать отдельные экстракты лекарственных трав или их комбинации. Обычно растительные экстракты представляют собой сложные смеси сотен основных и минорных соединений. Вторичные метаболиты растений, даже присутствуя на низких концентрациях, могут быть важны для качества, безопасности и эффективности конечного продукта. Из-за географических и сезонных изменений и условий окружающей среды химический состав экстрактов растений может сильно меняться. Кроме того, на содержание метаболитов могут влиять многие другие факторы, включая зрелость (период сбора урожая), процессы сушки и хранения, что в совокупности с вышеуказанными внешними факторами делает задачу классификации и определения состава таких объектов еще более сложной. Высокоинформативные хроматомасс-спектрометрические методы позволяют зарегистрировать сигналы разных по структуре компонентов растительных экстрактов и выявить значимые для проведения такой классификации маркеры. Однако применение таких современных методов анализа предполагает получение и последующую обработку больших объемов данных, что требует использования методов хемоинформатики и машинного обучения. Применение этих методов к так называемым «сырым» данным осложнено из-за присутствия шумовых сигналов и дрейфа базовой линии, и другими факторами, связанными с изменением условий измерения в ходе анализа большого числа образцов. По этой причине исходные данные обычно представляют в виде набора пиков, каждый из которых имеет свое значение времени удерживания и значение m/z основного сигнала, по которому этот пик был обнаружен. Для этого применяют разные алгоритмы фильтрации (и отсечения) шумовых сигналов, разметки, сглаживания и деконволюции пиков. Поскольку одному соединению на масс-хроматограмме может соответствовать несколько пиков (по сигналам изотопологов и аддуктных ионов), далее избыточные пики удаляют, формируют финальную таблицу признаков и определив молекулярные формулы соответствующих им соединений по полученному точному значению молекулярной массы и паттерну изотопного расщепления. Такой подход, используемый в ненаправленном хроматомасс-спектрометрическом анализе, обладает как явными преимуществами, так и некоторыми недостатками. Преимуществом является прямая интерпретируемость получаемых выводов, так как статистически отличающиеся признаки для заданной группы образцов могут быть идентифицированы как соединения-маркеры этой группы. Недостатками же являются потери химической информации при объединении неразрешенных пиков и игнорировании сигналов некоторых молекулярных и фрагментных ионов. Кроме того, такой подход не

применяют к данным, полученным методом высокоэффективной жидкостной хроматографии в сочетании с масс-спектрометрическим детектированием низкого разрешения, которые будут содержать гораздо больше перекрывающихся пиков по каждому значению m/z и также могут быть источником полезной информации. В этой связи, актуальным является создание алгоритмов анализа больших данных, основанных на разных способах предобработки и преобразования исходных ВЭЖХ-МС данных низкого и высокого разрешения с методами машинного обучения «с учителем» и «без учителя».

Цель работы заключалась в разработке аналитических подходов к проведению классификации растительных материалов и выявлению характеристических маркеров, на основе высокоэффективного хроматографического разделения с масс-спектрометрическим детектированием и методов машинного обучения.

Для достижения поставленной цели необходимо было решить следующие задачи:

- Оценить возможность использования данных целевого ВЭЖХ-МС анализа для идентификации и классификации растительных материалов по выбранным соединениям-маркерам.
- Выбрать условия получения и ненаправленного ВЭЖХ-МС анализа экстрактов из разных частей растений.
- Разработать способы предобработки получаемых трехмерных массивов данных для последующего применения методов машинного обучения.
- Выбрать параметры применяемых вычислительных методов с обучением «с учителем» и «без учителя» для решения задач классификации и кластеризации образцов, а также для поиска маркеров, ответственных за отнесение образца к какой-либо группе или кластеру.
- Проверить работоспособность предложенных подходов с использованием наборов образцов растений разных видов и разного происхождения.

Научная новизна

Предложены оригинальные варианты предобработки и преобразования данных ВЭЖХ-МС профилей низкого и высокого разрешения и продемонстрирована их применимость для последующего использования алгоритмов машинного обучения. Предложен быстрый способ кластеризации для трехмерных массивов ВЭЖХ-МС данных с помощью тензорного разложения по методу PARAFAC.

Проведено сравнение эффективности работы предложенных подходов к кластеризации образцов и выявлению значимых для классификации маркеров, основанных на развертке тензора ВЭЖХ-МС данных в сочетании с методами главных и независимых компонент, методом неотрицательного матричного разложения или

методом отбора признаков. Предложен алгоритм ранжирования по вариабельности содержания отдельных соединений в образцах.

Предложены новые способы предобработки и преобразования исходных данных ВЭЖХ-МС низкого разрешения, которые в сочетании с применением метода опорных векторов и сверточных нейронных сетей позволили классифицировать экстракты разных частей растений

Практическая значимость

Предложен быстрый способ экстракции и ВЭЖХ-МС анализа в широком диапазоне по значениям m/z и по содержанию органического растворителя в подвижной фазе, который позволяет регистрировать наиболее полный профиль разных по свойствам компонентов растительного сырья.

Предложены алгоритмы обработки исходных ВЭЖХ-МС данных низкого и высокого разрешения, включающие устранение шумов, сглаживание, выравнивание и изменение шага для шкалы времени и шкалы m/z, которые позволяют представить полученные данные в удобном для дальнейшего применения методов машинного обучения формате.

Разработанные программные алгоритмы для решения задач предобработки ВЭЖХ-МС данных, классификации образцов и поиска биомаркеров доступны исследователям в сети Интернет[1].

На основе ВЭЖХ-МС и метода PARAFAC предложен быстрой способ кластеризации образцов, позволяющий разбить исследуемые объекты на основные группы, а также выявить наиболее значимые хроматографические пики и m/z сигналы, характерные для образцов из выделенных групп.

В результате применения разработанных подходов выявлены и предварительно идентифицированы 23 потенциальных хемотаксономических маркера для разных видов растений из семейства Зонтичные, а также 8 маркеров, характерных для различных частей этих растений.

Положения, выносимые на защиту

1. Разработанная схема выбора уникальных характеристичных маркеров и маркеров качества с их последующим целевым ВЭЖХ-МС определением в режимах мониторинга выбранных ионов и ионных переходов позволяет проводить идентификацию лекарственных растений, находящихся в свободной продаже.

2. Разработанный на основе комбинации ВЭЖХ-МС анализа и тензорного разложения полученных массивов данных по методу PARAFAC подход позволяет разделять образцы, содержащие отдельные растительные материалы, их двойные и

[1] https://github.com/turovapolina/HPLC-MS-PARAFAC, https://github.com/turovapolina/unsupervised-LC-MS-data-treatment

тройные модельные смеси, а также экстракты из продуктов, содержащих эти материалы в качестве ароматизирующих добавок без использования индивидуальных стандартных соединений.

3. Предложенные способы предобработки и организации ВЭЖХ-МС данных низкого и высокого разрешения обуславливают успешность последующего применения методов машинного обучения «без учителя» для кластеризации экстрактов из листьев разных видов растений одного семейства (на примере семейства Зонтичных).

4. Комбинация извлечения водно-метанольной смесью и последующего разделения методом градиентной обращенно-фазовой ВЭЖХ в широком диапазоне концентраций органического растворителя с МС детектированием в режиме сканирования позволяет получать высокоинформативные хроматографические профили образцов растительных экстрактов, в которых были обнаруженны и идентифицированы компоненты, относящиеся к флавоноидам, кумаринам, гликозидам, липидам и хлорофиллам.

5. Разработанный подход на основе применения метода SVM к развернутому тензору предобработанных ВЭЖХ-МС данных низкого разрешения позволяет классифицировать образцы экстрактов из различных частей растений одного семейства (на примере семейства Зонтичных): листьев, стеблей, корней, плодов/соцветий.

6. Представление ВЭЖХ-МС данных в виде двумерных массивов с равной размерностью по осям с последующим применением сиамских нейронных сетей позволяет достигать сопоставимой с методом SVM точности классификации растительных экстрактов.

7. Использование предложенных способов аугментации ВЭЖХ-МС данных повышает точность классификации как в случае использования метода SVM, так и нейронных сетей.

Степень достоверности

Степень достоверности результатов проведенных исследований обеспечивалась применением современного хроматографического и масс-спектрометрического оборудования.

Соответствие паспорту научной специальности

Выпускная квалификационная работа соответствует паспорту специальности 02.00.02 – Аналитическая химия по областям исследований:

- методы химического анализа (химические, физико-химические, атомная и молекулярная спектроскопия, хроматография, рентгеновская спектроскопия, масс-спектрометрия, ядерно-физические методы и др);
- теория и практика пробоотбора и пробоподготовки в аналитической химии;
- математическое обеспечение химического анализа;
- анализ природных веществ.

Апробация результатов исследования

Основные результаты работы были представлены на конференциях:

2022 год: Международная конференция «13th Winter symposium on Chemometrics», Москва, Россия, 28 февраля – 4 марта;

2021 год: Международная конференция «ROAD TO CAC 2022 // ZOOM CONFERENCE ON 20-21 JULY 2021», Курмайёр, Италия, 20 – 21 июля; VI Всероссийская конференция с международным участием «Разделение и концентрирование в аналитической химии и радиохимии», Краснодар, Россия, 26 сентября – 2 октября; X Всероссийская конференция с международным участием «Масс-спектрометрия и ее прикладные проблемы», Москва, Россия, 18 – 22 октября;

2020 год: IV Всероссийская конференция с международным участием «Аналитическая хроматография и капиллярный электрофорез», Краснодар, Россия, 28 сентября – 2 октября; Международная конференция «4th International Symposium on Phytochemicals in Medicine and Food», Сиань, Китай, 30 ноября – 4 декабря;

2019 год: «48th International Symposium on High-Performance Liquid Phase Separations and Related Techniques», Милан, Италия, 16 – 20 июня; III Всероссийская конференция по аналитической спектроскопии с международным участием, Краснодар, Россия, 29 сентября – 5 октября; IX Всероссийская конференция с международным участием «Масс-спектрометрия и ее прикладные проблемы», Москва, Россия, 15 – 18 октября.

Публикации

По материалам работы опубликовано 12 печатных работ, в том числе три статьи в рецензируемых научных изданиях, индексируемых международными базами данных (Web of Science, Scopus) и рекомендованных в диссертационном совете МГУ по специальности 02.00.02 «Аналитическая химия», и 9 тезисов докладов на российских и международных конференциях.

Личный вклад автора

Личный вклад автора состоял в общей постановке задач, систематизации литературных данных, подготовке и проведении всех экспериментальных этапов исследования, обработке, интерпретации и оформлении полученных экспериментальных данных, подготовке материалов к публикации и представлении полученных результатов на конференциях. Все исследования, описанные в работе, выполнены лично автором или в сотрудничестве с коллегами.

Структура и объем работы

Полный текст работы состоит из 6 глав и включает 167 страниц, в том числе 44 рисунка, 12 таблиц. Список литературы содержит 179 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность выбранной темы, сформулированы цель исследования и поставлены задачи для достижения цели, показана научная новизна работы и ее практическая значимость.

Первая глава представляет собой обзор литературы, в котором систематизированы основные стадии и подходы к обработке масс-спектрометрических данных, включающие организацию данных, фильтрацию, хемометрический анализ.

В разделе «Организация (методы представления) данных» систематизирована информация об основных подходах к представлению трехмерных данных: создание таблиц с признаками – веществами; метаболомный подход, в котором генерируется таблица с основными пиками; работа с исходным тензором данных и развертка тензора в матрицу. Для каждого подхода приведены примеры из литературы.

В разделе «Фильтрация сигнала и сглаживание шума» обобщены основные подходы для подавления шума, усиления полезного сигнала, повышения разрешения. Основное внимание уделено методам непрерывного вейвлет-преобразования и фильтру Савитского-Голея и их применению для хроматографических и масс-спектрометрических данных.

Раздел «Методы обработки данных» является самым обширным и важным разделом литературного обзора. В него включено описание методов анализа данных «без учителя» и «с учителем». В подраздел «Методы без учителя» входит описание тензорных разложений (разложение Таккера и PARAFAC), их основные принципы и специфика применения для ВЭЖХ-МС и ГХ-МС данных. Далее обобщена информация о матричных методах факторизации, таких как PCA, ICA, NMF, HCA. Обсуждены варианты их применения к данных ВЭЖХ-МС, ГХ-МС, MSI для кластеризации данных, распознавания образцов, визуализации данных поиска потенциальных маркерных соединений, для обнаружения и нивелирования сигналов мешающих веществ. В подразделе «Методы с учителем» описаны основные подходы для классификации, которые применяют к хроматографическим и масс-спектрометрическим данным. В подраздел вошли такие методы, как PLS-DA, SVM, нейронные сети (полносвязные и сверточные). Детально показано устройство этих методов, а также результаты их применения для выявления различий между образцами, классификации, извлечения значимых сигналов из массива образцов.

На основании обзора литературы сделаны выводы, которые подтверждают актуальность выбранной темы исследования и способов решения поставленных задач.

Вторая глава включает описание используемых в работе химических реактивов, материалов и оборудования; условий и техники проведения экспериментов, а также обработки данных.

В работе использовали следующее аналитическое оборудование:

ВЭЖХ-МС систему, состоящую из гибридного тандемного квадрупольного масс-спектрометрического детектора AB Sciex Qtrap 3200 (Канада) с линейной ионной ловушкой, оснащенного источником электрораспылительной ионизации; и жидкостного хроматографа Dionex Ultimate 3000 (США). Регистрацию хроматограмм и обработку данных проводили при помощи программного обеспечения Analyst 1.5.1 (Канада).

ВЭЖХ-МСВР систему, состоящую из масс-спектрометрического детектора с орбитальной ионной ловушкой Orbitrap Exactive (Германия), оснащенного источником нагреваемой электрораспылительной ионизации; и жидкостного хроматографа Thermo Scientific Acella HPLC system (США). Регистрацию хроматограмм и обработку данных проводили при помощи программного обеспечения Xcalibur™ Software (версия 2.2) предоставленную компанией Thermo Scientific™.

ВЭЖХ-МСВР систему, состоящую из времяпролетного масс-спектрометрического детектора Bruker Impact II high-resolution Quadrupole Time-of-Flight (Германия) и жидкостного хроматографа Bruker Elute LC system (Германия).

Разделение проводили на следующих хроматографических колонках: Acclaim RSLC C18 120 Å (150 × 2.1 мм), диаметр зерна сорбента 3 мкм (Thermo Scientific™, США) и Intensity Solo C18 90 Å (100 × 2.1 mm) диаметр зерна сорбента 1.8 мкм (Bruker, Германия).

Третья глава («Идентификация и классификация растительных материалов на основе данных целевого ВЭЖХ-МС анализа») посвящена оценке возможностей целевого ВЭЖХ-МС анализа для идентификации коммерчески доступных растительных материалов. Предложенная схема выбора и обнаружения биомаркеров методом ВЭЖХ-МС может быть использована для создания внутренней «*in-house*» библиотеки (таблицы, содержащей информацию о целевых соединениях и сигналам для их обнаружения) для целевого анализа, которая поможет определить биомаркеры в экстрактах продаваемых на рынке продуктов, идентифицировать растения, используемые в их производстве, осуществлять первичный контроль качества. В зависимости от наличия и числа биомаркеров в экстракте растения предлагается три метода работы с образцом. Если для рассматриваемого растения существуют уникальные соответствующие ему биомаркеры, то есть два возможных пути работы с такими экстрактами. Первый способ используется, когда найдено небольшое число уникальных маркеров. В этом случае в таблицу целевых сигналов добавляют по два наиболее интенсивных ионных перехода (реакции) для каждого уникального маркера, которые могут быть взяты из МС/МС спектров каждого выбранного биомаркера. Второй путь используется, если в химическом составе растительного экстракта есть группа биомаркеров. В таком случае можно проводить анализ в режиме мониторинга выбранных характеристичных ионов. Эти характеристичные ионы в масс-спектрах

должны соответствовать общему фрагменту (фрагментам) структуры этой группы биомаркеров. В случае, если для исследуемого растения нет характеристических биомаркеров, то контролируется соотношение концентраций нескольких целевых соединений, служащих маркерами качества (флавоноиды, кислоты и т. д.). Содержания флавоноидных маркеров качества в нескольких исследованных растениях приведены в Таблице 1.

Таблица 1. Содержание маркеров качества в растительных материалах

Растение	Концентрация маркеров качества, мкг/г			
	Хризоериол	Акацетин	Изоакацетин	Диосметин
Ocimum basilicum	0.41	0.68	2.24	1.50
Mentha arvensis	<0.1	<0.1	20.7	5.9
Origanum vulgare	0.83	3.0	14.5	3.2
Rosmarinus officinalis	2.04	16.7	161	3.3
Hypericum perforatum	0.65	<0.1	-	<0.2
Chrysanthemum morifolium	<0.1	23.2	-	3.4
	Хризин	Лютеолин-7-глюкоронид	Кверцетин-7-глюкозид	Кверцетин-3-глюкозид
Geranium pratense	124	2.5	113.8	135.3
Matricaria chamomilla	33	<0.07	75.7	38.5
Eryngium campestre	<0.02	0.87	51.3	198.5

На основе идентификации 61 выбранного биомаркера, в целях проведения быстрого и эффективного скрининга растительных экстрактов из 39 разных растений были предложены рекомендации для целевого ВЭЖХ-МС контроля качества растительных продуктов и пищевых добавок на их основе. Условия ультразвуковой экстракции, обеспечивающие высокое извлечение, в сочетании с селективным ВЭЖХ-МС анализом позволили достичь низких пределов обнаружения при целевом скрининге биомаркеров. С использованием индивидуальных стандартов, были выбраны условия детектирования всех выбранных биомаркеров. В результате была предложена схема выбора биомаркеров и режима детектирования в зависимости от их наличия в экстракте растения. Эти методы можно использовать для проверки подлинности и количественного измерения биологически активных соединений, но информативность такого подхода ограничена.

В **главе 4** («ВЭЖХ-МС-PARAFAC подход для идентификации растительных экстрактов») описаны анализ 34 образцов экстрактов растений, обработка массива,

полученного после их анализа, и интерпретация результатов. Идея предложенного подхода заключалась в том, чтобы сохранить как можно больше информации за счет сохранения структуры ВЭЖХ-МС данных, хотя для приведения матриц образцов к единой форме потребовались некоторые преобразования исходного трехмерного набора данных. Окончательный набор данных представлял собой тензор третьего порядка и имел размерности 34×400×1200. В результате PARAFAC разложения тензора получены 3 матрицы, при этом количество компонент PARAFAC было выбрано равным 3, что соответствует числу видов растений в использованном наборе образцов. Первая матрица с размерностью 34×3 является матрицей счетов, две другие – матрицей нагрузок: первая представляет времена (400×3), а вторая представляет отношение массы к заряду (1200×3). На Рис. 1 представлена визуализация счетов и разделение образцов на кластеры.

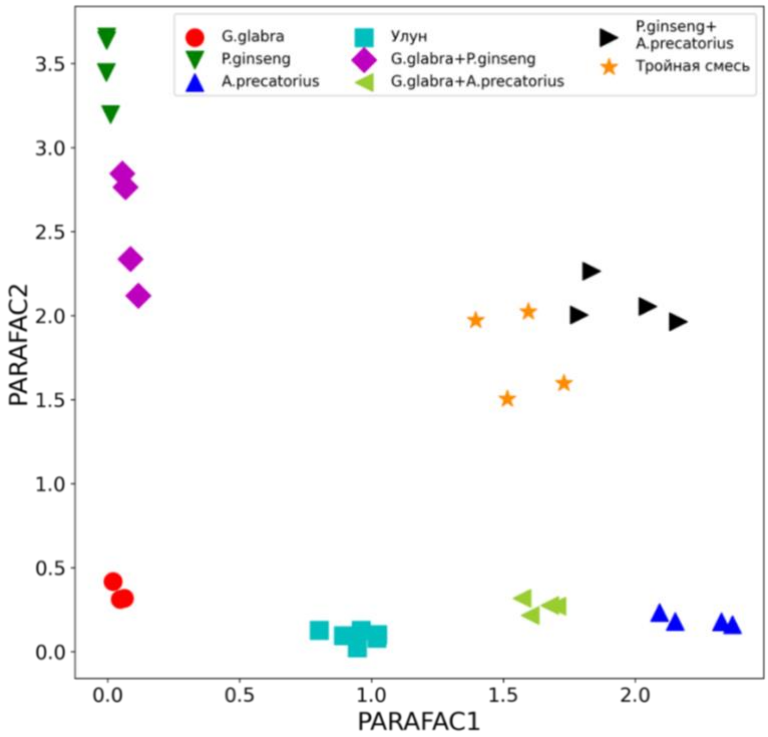

Рис. 1. Применение метода PARAFAC к подготовленному трехмерному набору ВЭЖХ-МС данных после сглаживания и вычитания шума. Визуализация счетов PARAFAC первой и второй компоненты.

Разделение групп было подтверждено применением метода k-средних к матрице счетов и иерархическим кластерным анализом. При применении метода k-средних было задано исходное количество классов равное 8, после его применения все образцы были помечены правильными классами (индекс Рэнда составил 100 %). Кластерный анализ проводился с использованием евклидовой метрики для вычисления расстояния между точками и взвешенного попарно группового метода расчета расстояния между кластерами. На дендрограмме, построенной по результатам иерархичного кластерного анализа, два образца из класса тройных смесей были отнесены к группе двойной смеси *P. ginseng* и *A. precatorius*, поскольку эти две группы были недостаточно разделены.

Таким образом, предложенный способ анализа данных ВЭЖХ-МС дал возможность осуществить оценку присутствия экстрактов в смеси. В частности, это может быть использовано для идентификации и распознавания сладких на вкус растительных материалов (листья *A. precatorius*, корни *G. glabra* и *P. ginseng*) в различных традиционных лекарственных препаратах и пищевых продуктах, что и было продемонстрировано (Рис. 1) на примере выделения группы образцов ароматизированного чая (улуна).

Для интерпретации результатов нагрузки, 1-ой и 2-ой компонент PARAFAC (PARAFAC1 и PARAFAC2), отвечающие за времена удерживания, сравнивали с масс-хроматограммами чистых экстрактов по выделенным ионам, созданными суммированием интенсивностей ионов фрагментов сапогенинов (m/z 485, 467, 449, 439, 421 для *A. precatorius* и m / z 443, 441, 425, 423, 407 для *P. ginseng*). Для образцов *A. precatorius* хроматограммы по выделенным ионам для ионов абрусогенина имеют ту же форму, что и временные нагрузки от PARAFAC1 (см. Рис. 2 А, В)). Для образцов *P. ginseng* такие же хроматограммы для фрагментных ионов с протопанаксатриольным и протопанаксадиольным остовом аналогичны нагрузкам из PARAFAC2 (см. Рис. 2 Б, Г).

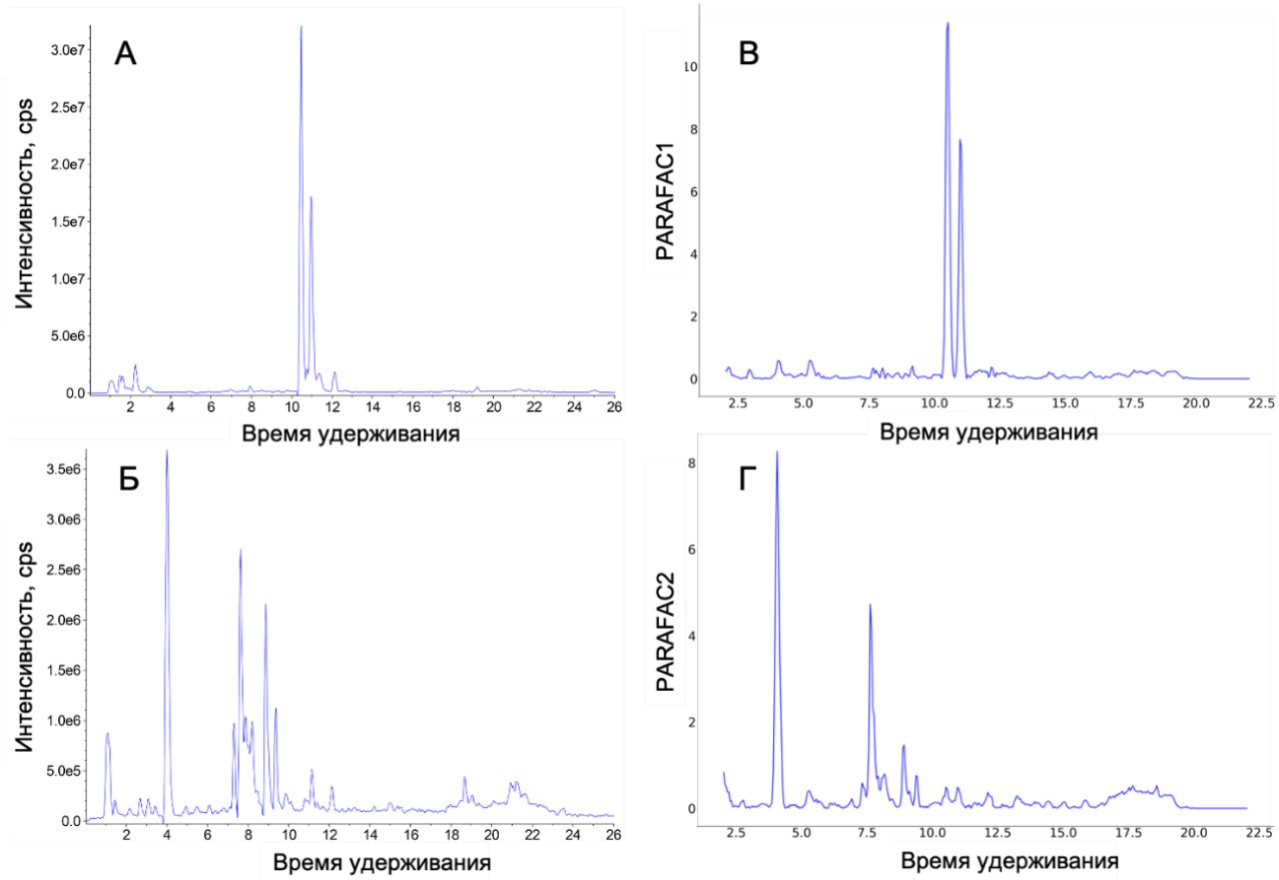

Рис. 2. А - суммарный сигнал по выделенным ионам с m/z 485, 467, 449, 439, 421 в образце *A. precatorius*; Б - суммарный сигнал по выделенным ионам m/z 443, 441, 425, 423, 407 в образце *P. ginseng*; В – нагрузки по времени компоненты PARAFAC1; Г - нагрузки по времени компоненты PARAFAC2.

Далее интерпретировали нагрузки по массе, ранжировали значения m/z по наибольшим весами из PARAFAC1 и PARAFAC2. Можно заметить, что все ионы из компоненты PARAFAC1 соотносятся с фрагментными характеристическими ионами гинсенозидов из экстрактов корней *P. ginseng* (Таблица 2). Эти сигналы включают сигналы ионов, полученных последовательным отщеплением 1-2 сахарных фрагментов и молекул H_2O. Кроме того, все ионы из PARAFAC2 хорошо соотносятся с характеристическими фрагментными ионами абрусогенина. Таким образом, можно заключить, что PARAFAC1 отвечает за биомаркеры, связанные с абрусом а PARAFAC2 за биомаркеры женьшеня, и разработанный подход позволил обнаружить эти зависимости в относительно широком диапазоне масс и времен удерживания.

Таблица 2. Значения m/z с наибольшим весом компонент PARAFAC1 и PARAFAC2, где $C_{30}H_{44}O_5$ - абрусогенин (А), $C_{30}H_{52}O_3$ - агликон протопанаксадиола (ППД), $C_{30}H_{52}O_4$ - агликон протопанаксатриола (ППТ).

PARAFAC$_1$			PARAFAC$_2$		
m/z	Вес	Интерпретация	m/z	Вес	Интерпретация
450	3.62	А $[^{13}CC_{29}H_{44}O_5–2H_2O+H]^+$	443	3.75	ППД $[C_{30}H_{52}O_3–H_2O+H]^+$
449	3.61	А $[C_{30}H_{44}O_5–2H_2O+H]^+$	570	3.67	ППД $[^{13}CC_{29}H_{52}O_3–3H_2O+Glc+H]^+$
440	3.58	А $[^{13}CC_{29}H_{44}O_5–H_2O–CO+H]^+$	569	3.47	ППД $[C_{30}H_{52}O_3–3H_2O+Glc+H]^+$
439	3.58	А $[C_{30}H_{44}O_5–H_2O–CO+H]^+$	442	3.46	ППТ $[^{13}CC_{29}H_{52}O_4 – 2H_2O+H]^+$
404	3.55	А $[^{13}CC_{29}H_{44}O_5–3H_2O–CO+H]^+$	424	3.46	ППТ $[^{13}CC_{29}H_{52}O_4 – 3H_2O+H]^+$
421	3.52	А $[C_{30}H_{44}O_5–2H_2O–CO+H]^+$	405	3.45	ППТ $[C_{30}H_{52}O_4 – 4H_2O + H]^+$
468	3.48	А $[^{13}CC_{29}H_{44}O_5–H_2O+H]^+$	588	3.43	ППД $[^{13}CC_{29}H_{52}O_3 – 2H_2O+Glc+H]^+$
422	3.46	А $[^{13}CC_{29}H_{44}O_5–2H_2O–CO+H]^+$	587	3.43	ППД $[C_{30}H_{52}O_3 – 2H_2O+Glc+H]^+$
469	3.41	А $[^{13}C_2C_{28}H_{44}O_5–H_2O+H]^+$	406	3.40	ППТ $[^{13}CC_{29}H_{52}O_4 – 4H_2O+H]^+$
441	3.35	А $[^{13}C_2C_{28}H_{44}O_5–H_2O–CO+H]^+$	622	3.37	ППТ $[^{13}CC_{29}H_{52}O_4–H_2O+Glc+H]^+$
485	3.33	А $[C_{30}H_{44}O_5+H]^+$	790	3.35	ППТ $[C_{30}H_{52}O_4 +Glc+Rha–H_2O+Na]^+$ ППД $[C_{30}H_{52}O_3+2Glc–H_2O+Na]^+$
467	3.32	А $[C_{30}H_{44}O_5–H_2O+H]^+$	425	3.33	ППД $[C_{30}H_{52}O_3 – 2H_2O+H]^+$
487	3.29	А $[^{13}C_2C_{28}H_{44}O_5+H]^+$	441	3.33	ППТ $[C_{30}H_{52}O_4 – 2H_2O+H]^+$

Таким образом, предлагаемый ВЭЖХ-МС-PARAFAC метод может быть использован для нецелевого скрининга и контроля качества продуктов на основе растительного сырья без использования индивидуальных стандартных образцов. Этот метод основан на анализе «сырых» массивов данных, полученных в результате ВЭЖХ-МС анализа, и не требует развертки тензора данных, поэтому в нем минимальна потеря информации. Наличие определенного растительного материала в образцах каждого из сформированных кластеров может быть подтверждено при интерпретации матриц нагрузок и построении характеристичных хроматограмм.

В **главе 5** («Методы обработки ВЭЖХ-МС данных «без учителя» и их применение для поиска потенциальных хемотаксономических маркеров») для данных ВЭЖХ-МСНР и ВЭЖХ-МСВР анализа экстрактов из листьев 19 растений семейства Зонтичных были рассмотрены два варианта обработки: тензорное разложение с помощью PARAFAC и матричное разложение после развертки тензора. Для развернутых тензоров были рассмотрены четыре подхода для уменьшения размерности и факторизации данных: PCA, ICA, NMF и UFS. Проведено сравнение результатов, полученных этими методами для обоих наборов данных по различным критериям.

Все образцы модельного набора были проанализированы в одинаковых хроматографических условиях с помощью двух хроматомасс-спектрометров высокого и низкого разрешения. Градиентная программа хроматографического разделения была задана таким образом, чтобы состав подвижной фазы менялся в широком диапазоне. На стадии предобработки данных был использован метод линейной интерполяции для унификации шкалы времен. Шкала масс также была унифицирована как для высокого, так и для низкого разрешения.

Первым было применено разложение PARAFAC, число компонент подбирали с помощью процедуры многократного разделения массива на две части и последующего независимого разложения этих частей. Затем проводилась процедура ассоциирования компонент разложений между собой и поиск коэффициентов корреляции Такера (TCC). На Рис. 3 показано среднее значение TCC для моделей с разным числом компонент. Было принято оптимальным число компонент равное 6, поскольку медианное значение TCC для всех разбиений при таком числе компонент оказывается одним из самых высоких, а дисперсия значений самая низкая.

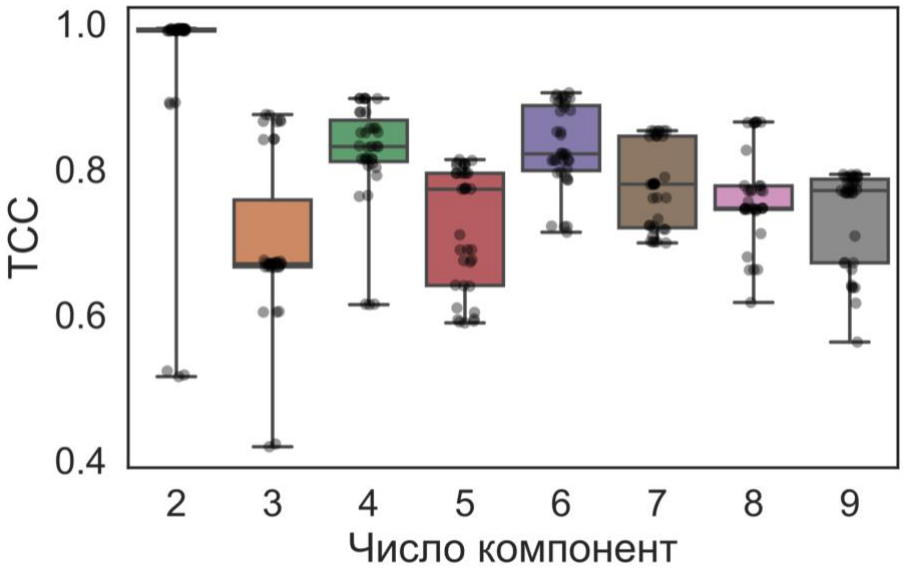

Рис. 3. Зависимость коэффициента корреляции Такера от числа компонент.

В качестве предподготовки для методов матричной факторизации была произведена развертка тензора таким образом, чтобы количество образцов оставалось

неизменным, а два других измерения (в случае ВЭЖХ-МС данных это m/z и время удерживания) объединялись в одно новое измерение. Далее, к полученной матрице были применены методы PCA, ICA, NMF, UFS. Для выбора числа компонент для каждого из методов были предложены следующие подходы:

1) PCA: оптимальным числом компонент то, которое будет достаточным для объяснения 95% дисперсии данных;
2) ICA: метод ICA-by-block. Данные были разделены на 2 блока, для каждого блока были рассчитаны модели ICA с числом компонент от 1 до 10. Затем модели двух блоков, рассчитанные с одинаковым числом компонент, сравнивали путем вычисления корреляции между каждой парой нагрузок. На Рис. 4 А, Б представлены графики корреляции сигналов двух наборов данных. Для обоих типов данных после извлечения более 4 компонент график начинает падать, что означает, что корреляции между компонентами различных блоков становится ниже. Таким образом, оптимальное число компонент в этих наборах данных - 4;
3) NMF: оптимальное количество компонент определяется таким, чтобы на графике остаточная сумма квадратов (RSS)-число компонент наблюдалась точка перегиба (Рис. 4 В, Г);
4) UFS: исключены признаки с дисперсией ниже заранее определенного порога, который в данном случае был задан как среднее значение дисперсии для всех признаков.

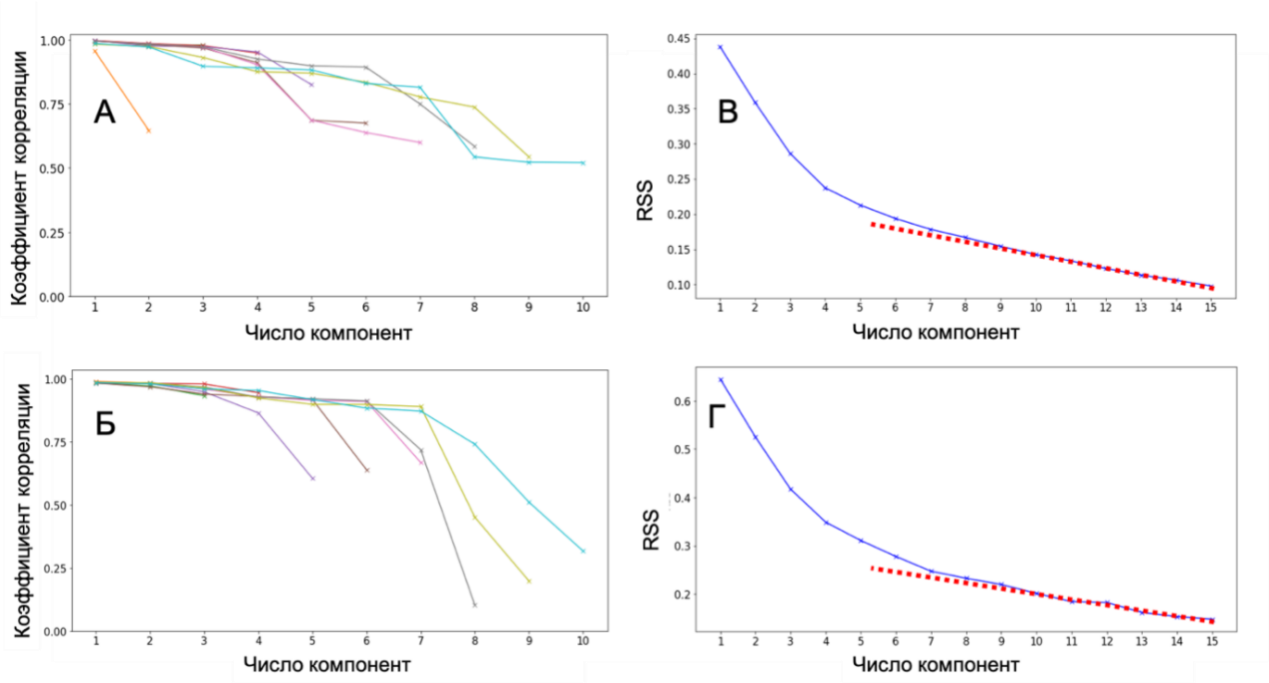

Рис. 4. Процесс определения количества компонент ICA для данных ВЭЖХ-МСНР (А) и ВЭЖХ-МСВР (Б); компонентов NMF для данных ВЭЖХ-МСНР (В) и ВЭЖХ-МСВР (Г). Красные линии на рисунках (В) и (Г) представляют собой линейную аппроксимацию методом наименьших квадратов последних шести точек графика.

Результаты применения алгоритмов сравнивались с точки зрения нескольких критериев: индекс Дэвиса-Болдина, значение критерия силуэта, время вычисления и количество шумовых компонент. Результаты представлены в Таблице 3.

На основе рассмотренных критериев методы PCA, ICA и UFS демонстрируют лучшие результаты и могут считаться наиболее подходящими методами для обработки ВЭЖХ-МСНР и ВЭЖХ-МСВР данных в случае их представления в виде развернутого тензора.

Таблица 3. Сравнение методов обработки данных

Метод	**Индекс Дэвиса-Болдина**	**Критерий силуэта**	**Время вычисления, сек**	**Шумовые компоненты**
ВЭЖХ-МСНР данные				
PCA	0.32	0.71	4.56	1
ICA	0.47	0.63	5.71	1
NMF	0.50	0.60	104.66	3
PARAFAC	0.52	0.51	151.27	2
UFS	0.52	0.52	1.44	–
ВЭЖХ-МСВР данные				
PCA	0.83	0.48	2.48	0
ICA	1.25	0.44	1.80	0
NMF	1.90	0.25	78.42	1
PARAFAC	1.05	0.38	122.86	1
UFS	0.75	0.40	0.26	–

Далее был предложен алгоритм отбора признаков, которые потенциально могут относиться к хемотаксономическим маркерам. Было решено извлечь признаки, соответствующие 50 соединениям из результатов трех методов, показавших более высокие результаты. Далее было получено пересечение всех списков этих сигналов. Примерно 50% сигналов из списков, полученных из каждого метода, были зафиксированы в списке пересечений. Из них были предварительно идентифицированы 23 соединения путем сравнения их спектров МС/МС с литературой и имеющимися базами данных. Распределение биомаркеров в образцах приведено на Рис. 5.

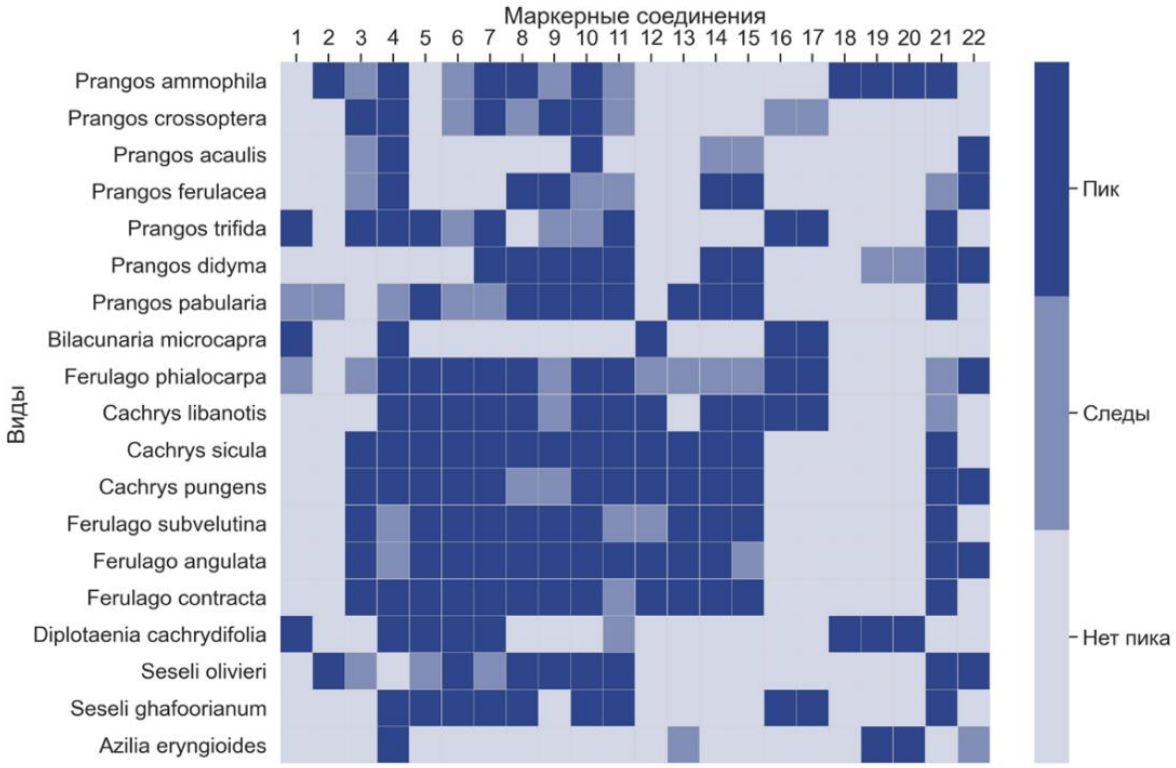

Рис. 5. Распределение распознанных биомаркеров в исследуемых образцах.

Наконец, для результатов, полученных каждым методом, были построены филогенетические деревья. Следует отметить, что деревья, полученные на основе ВЭЖХ-МС данных, демонстрируют различия в химическом составе, которые не коррелируют с результатами молекулярно-филогенетического анализа растений. Для этой задачи был использован подход, который включает в себя вычисление попарных расстояний между всеми образцами и построение матриц расстояний с добавлением в качестве количественной характеристики поточечно рассчитанной среднеквадратичной ошибки (MSE), которая может быть рассчитана по формуле (1) (где $I^1_{i,j}$ и $I^2_{i,j}$ - элементы i, j первой и второй матрицы расстояний соответственно) а вместо значений пикселей исходные значения расстояния в матрицах.

$$MSE\ (I^1, I^2) = \frac{1}{n}\sum_{i,j}(I^1_{i,j}\ -\ I^2_{i,j})^2 \qquad (1)$$

В результате сравнения ошибки, вычисленных с использованием предложенного подхода, оказалось, что наименьшее значение было получено методом UFS для данных ВЭЖХ-МСНР и ВЭЖХ-МСНР: 0.105 и 0.144 соответственно.

В **главе 6** («Методы обработки ВЭЖХ-МС данных с учителем и их применение для поиска маркеров») было показано применение к данным ВЭЖХ-МС методов обучения «с учителем». Расширенный набор образцов использованных в этой части исследования состоял из 186 образцов растительных экстрактов из разных частей (корни, стебли, листья, плоды/цветы) 19 видов растений, представляющих 7 родов семейства Зонтичные. Таким образом, стала возможной классификация образцов и по

частям, и по родам растений. Образцы были проанализированы методом ВЭЖХ-МСНР в выбранных ранее условиях и проведена предобработка, описанная в главе 5.

Первым методом, выбранным для классификации образцов, стал метод SVM, примененный к развернутому тензору данных. Для того, чтобы из можно было выделить признаки, характерные для каждого класса, дополнительно использовали линейную функцию ядра. Ввиду большого разброса химического состава внутри классов растений, принадлежащих к одному роду и сильного дисбаланса классов точность классификации по родам была довольно низкой и было решено классифицировать принадлежность образцов к разным частям растений (корни, стебли, листья, плоды/соцветия).

После анализа набора данных методом SVM был предложен алгоритм извлечения сигналов, относящихся к каждому из классов. Для этого была выделена матрица весов для каждого из признаков по каждому классу. Признаки были отранжированы по весу и из них отобраны первые 2000 с максимальным весом. Далее каждый признак из каждого набора был проверен на появление пика данного сигнала в образцах, не принадлежащих к изучаемому классу. Таким образом, удалось выявить 8 характеристичных маркеров для 4 заданных классов образцов. Примеры хроматограмм маркеров корней и стеблей представлены на хроматограммах на Рис. 6. Обнаруженные маркеры корней относились к дигликозидам кумаринов, ранее выделенным из корней родственных растений, а маркерами плодов оказались фосфадитилхолиновые липиды. Характеристичными сигналами для стеблей и листьев оказались фрагментные ионы хлорофиллов.

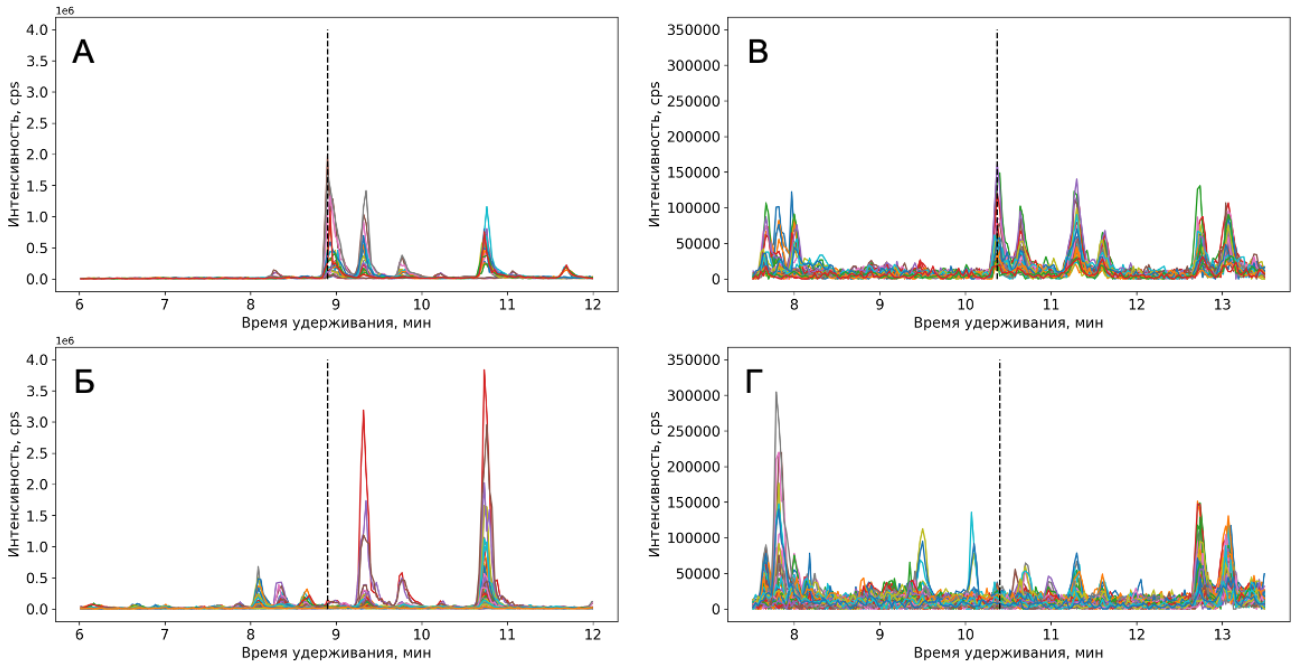

Рис. 6. Наложение масс-хроматограмм образцов (А) корней, содержащих маркер 1 (8.96 мин, m/z 307); (Б) всех остальных образцов не содержащих маркер 1; (В) стеблей, содержащих маркер 3 (10.37 мин, m/z 607); (Г) всех остальных образцов не содержащих маркер 3.

На следующем шаге исследования была изучена возможность применения сверточных нейронных сетей в обработке ВЭЖХ-МС данных. Для работы с

нейронными сетями данные были приведены к одинаковой размерности (750×750) по обеим осям и нормализованы. В качестве архитектуры была взята сеть ResNet18 с кросс-энтропией в качестве функции потерь. Работа сети была оценена с помощью кросс-валидации (разбиение на 3 части). Для процесса обучения и составления батча в нейронной сети был использован взвешенный сэмплер, чтобы скомпенсировать дисбаланс классов.

Также, было предложено использовать сиамскую сеть с триплетной функцией потерь. Такие сети хорошо зарекомендовали себя в решении задач, в которых число представителей каждого класса мало или разница между классами очень невелика. За основу архитектуры данной сети также была взята сеть ResNet18. В использованной сети веса предобученны на находящемся в свободном доступе наборе данных ImageNet. Размерность последнего полносвязного слоя была изменена с 1000 до 32.

Для увеличения точности работы сетей были предложены способы аугментации. Для имеющихся данных были предложены алгоритмы добавления шума, растяжения и сглаживания по времени, имитирующие процессы, которые действительно могут происходить при небольшом изменении условий хроматомасс-спектрометрического анализа. Пример применения этих методов представлен на Рис. 7.

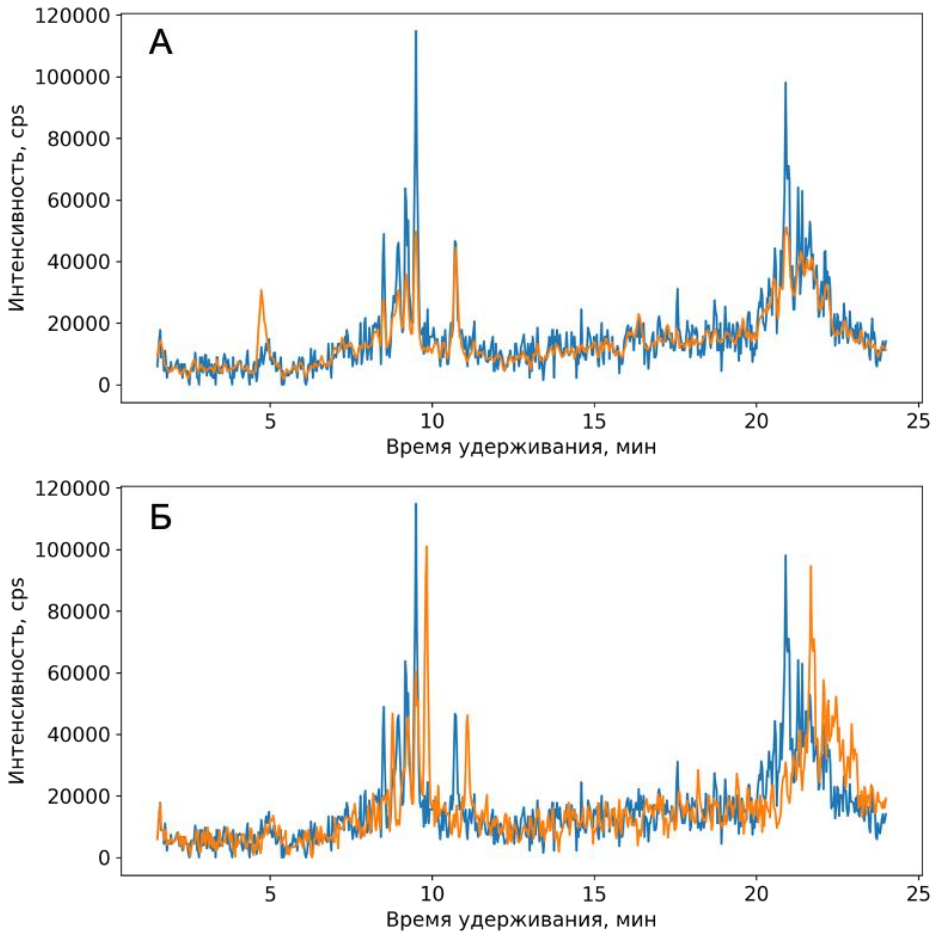

Рис. 7. Методы аугментации: (А) размывание с помощью нормализованного блочного фильтра, (Б) растяжение масс-хроматограммы.

При применении аугментации точность классификации для сети с классической архитектурой увеличилась с 78.6 до 98.7, а для сиамских сетей с 93.1 до 99.2 соответственно.

В **заключении** описаны основные результаты, полученные в работе.

Исследования, описанные в главе 3, посвящены оценке возможностей целевого ВЭЖХ-МС анализа для идентификации растений. Показано, что основании литературных данных могут быть выбраны индивидуальные и групповые характеристичные маркеры, а также маркеры качества, которые могут быть обнаружены в разных растительных образцах.

Глава 4 посвящена разработке ВЭЖХ-МС-PARAFAC подхода для быстрой кластеризации образцов. По данным, полученным из матрицы счетов, были выделены 8 отдельных кластеров образцов. Показано, что нагрузки, отдельных компонент, полученные методом PARAFAC и хроматограммы по выделенным ионам характеристических маркеров абруса и женьшеня имеют схожие профили. Также, среди сигналов, имеющих наибольший вес в матрице нагрузок по m/z, были обнаружены сигналы фрагментных ионов для тех же групповых характеристичных маркеров.

В главе 5 были рассмотрены основные аспекты применения методов машинного обучения «без учителя» (PCA, ICA, UFS, NMF) к предобработанным ВЭЖХ-МС данным высокого и низкого разрешения, представленным в форме развернутого тензора. Результаты разложений сравнивались с результатом применения метода PARAFAC к неразвернутому тензору. Для данных высокого разрешения была предложена схема унификации шкалы масс для всех образцов. На основе нескольких критериев методы PCA, ICA и UFS демонстрируют лучшие результаты как для данных МСНР, так и для МСВР. Кроме того, был разработан алгоритм выбора сигналов наиболее значимых маркерных соединений.

В главе 6 продемонстрирована применимость метода SVM к развернутому набору данных ВЭЖХ-МС низкого разрешения для классификации образцов экстрактов растений семейства Зонтичные (*f1*=84.4%). Выявлены характеристичные маркеры этих групп. Кроме того, была продемонстрирована возможность использования сверточных классических (*f1*=78.6%) и сиамских (*f1*=93.1%) нейронных сетей для работы с исходными данными ВЭЖХ-МС. Также было предложено 4 варианта аугментации для данных ВЭЖХ-МС и показано, что при их использовании точность классификации увеличивается.

ВЫВОДЫ

1. Предложена схема выбора биомаркеров для их последующего целевого ВЭЖХ-МС определения в режимах мониторинга выбранных реакций или выделенных ионов с целью идентификации растительных материалов. Выбраны условия обнаружения 61 биомаркера для 39 видов растений. Показано, что соотношение

концентраций некоторых распространенных маркеров в экстрактах растений может отличаться, что может быть использовано для их идентификации в отсутствие характеристичных маркеров. Показано, что в отсутствие характеристических маркеров могут быть использованы соотношения содержаний нескольких распространенных соединений, в частности флавоноидов.

2. Предложен подход для кластеризации образцов материалов лекарственных растений на основе комбинации хроматомасс-спектрометрического анализа их водно-метанольных экстрактов и тензорного разложения полученных массивов данных методом PARAFAC. Предложенный подход не требует применения индивидуальных стандартных соединений и использует данные масс-спектрометрии низкого разрешения. Показано, что нагрузки, полученные методом PARAFAC и хроматограммы по выделенным ионам характеристических маркеров имеют схожие профили.

3. Предложены подходы для работы с исходными ВЭЖХ-МС данными с использованием развертки тензора в матрицу и последующим применением методов PCA, ICA, NMF, UFS. На основе четырех критериев методы PCA, ICA и UFS превосходят методы PARAFAC и NMF как для данных МСНР, так и для МСВР, в частности, по критерию силуэта лучшим методом был PCA со значениями 0.71 и 0.48 для данных ВЭЖХ-МСНР и ВЭЖХ-МСВР соответственно.

4. Предложенные подходы были опробованы на наборе образцов растений разных видов и разного происхождения. Было показано, что наиболее значимые признаки, выделяемые с помощью предложенных алгоритмов, были одинаковыми и соответствовали характеристичным маркерным соединениям, которые были предварительно идентифицированы с помощью ВЭЖХ-МС/МС и ВЭЖХ-МСВР данных. Для растений семейства Зонтичные выделено 23 потенциальных хемотаксономических маркера, большей частью относящихся к классу кумаринов.

5. Предложен подход к классификации экстрактов из различных частей растений на основе применения метода SVM к развернутому тензору предобработанных ВЭЖХ-МС данных низкого разрешения. Точность предложенного подхода для набора образцов, состоящего из растений семейства Зонтинчные составила 84.4 %. Предварительно идентифицировано 8 выделенных характеристичных маркеров, соответствующих 4 группам исследованных образцов, а именно: корням, стеблям, листьям и плодам (соцветиям).

6. Продемонстрирована применимость сверточных нейронных сетей для работы с исходными («сырыми») данными ВЭЖХ-МС анализа. Точность сиамских сетей (93.1 %) оказалась несколько выше точности сети с классической архитектурой (78.6 %), а также выше точности метода SVM (84.4 %). Также было предложено 4 варианта аугментации для данных ВЭЖХ-МС и показано, что при их использовании точность.

Основные результаты работы изложены в следующих публикациях:

Научные статьи, опубликованные в рецензируемых научных журналах, индексируемых в базах данных Web of Science, Scopus, RSCI и рекомендованных для защиты в диссертационном совете МГУ по специальности 02.00.02 – «Аналитическая химия»:

1. Turova P., Styles I., Timashev V., Kravets K., Grechnikov A., Lyskov D., Samigullin T., Podolskiy I., Shpigun O., Stavrianidi A. Unsupervised methods in LC-MS data treatment: Application for potential chemotaxonomic markers search // J. Pharm. Biomed. Anal. 2021. Vol. 206. P. 114382. Импакт-фактор Web of Science – 3.935, **Q1**. (50 %)
2. Turova P., Rodin I., Shpigun O., Stavrianidi A. A new PARAFAC-based algorithm for HPLC-MS data treatment: Herbal extracts identification // Phytochem. Analysis. 2020. V. 31(6). P. 948–956. Импакт-фактор Web of Science – 3.373, **Q1**. (70 %)
3. Turova P., Stekolshchikova E., Baygildiev T., Shpigun O., Rodin I., Stavrianidi A. Unified strategy for HPLC-MS evaluation of bioactive compounds for quality control of herbal products. // Biomed. Chromatogr. 2018. V. 32(12). P. e4363. Импакт-фактор Web of Science – 1.902, **Q3**. (40 %)

Иные публикации:

4. Турова П.Н., Коряковцев П.А., Родин И.А., Ставрианиди А.Н. Разработка новых способов обработки массивов данных масс-спектрометрического анализа экстрактов из растительного сырья // Материалы III Всероссийской конференции по аналитической спектроскопии с международным участием. Краснодар, Россия. 2019. С. 167.
5. Турова П.Н., Ставрианиди А.Н. Применение ВЭЖХ-МС и метода PARAFAC для дискриминации образцов растительных материалов // Материалы IX Всероссийской конференции с международным участием «Масс-спектрометрия и ее прикладные проблемы». Москва, Россия. 2019. С. 6.
6. Stavrianidi A., Turova P., Rodin I. Development of new approaches for determination and identification of components from plant materials using HPLC-MS data // 48th International Symposium on High-Performance Liquid Phase Separations and Related Techniques. Milan, Italy. 2019. P. 204.
7. Турова П.Н., Голубева А.А., Тимашев В.И., Кравец К.Ю., Гречников А.А., Лысков Д.Ф., Шпигун О.А., Ставрианиди А.Н. Классификация образцов растений семейства umbelliferae по данным ВЭЖХ-МС анализа // Материалы IV Всероссийской конференции с международным участием «Аналитическая хроматография и капиллярный электрофорез». Краснодар, Россия. 2020. С. 47.
8. Turova P., Rodin I., Shpigun O., Stavrianidi A. Application of HPLC-MS-PARAFAC approach in phytochemical analysis // 4th International Symposium on Phytochemicals in Medicine and Food 4-ISPMF. Сиань, Китай. 2020. P. 145.
9. Турова П.Н., Ставрианиди А.Н. Применение методов обучения без учителя к

данным ВЭЖХ-МС анализа для поиска хемотаксонометрических маркеров // Материалы VI Всероссийского симпозиума «Разделение и концентрирование в аналитической химии и радиохимии» с международным участием. Краснодар, Россия. 2021. С. 173.

10. Turova P., Styles I., Lyskov D., Samigullin T., Shpigun O., Stavrianidi A. Exploration of unsupervised machine learning techniques possibilities in LC-MS data treatment // «ROAD TO CAC 2022. ZOOM CONFERENCE ON 20-21 JULY 2021». 2021. Курмайёр, Италия.

11. Турова П.Н., Ставрианиди А.Н. Новые подходы к обработке и анализу ВЭЖХ-МС данных, полученных при исследовании растительных экстрактов. // Материалы IX Всероссийской конференции с международным участием «Масс-спектрометрия и ее прикладные проблемы». Москва, Россия. 2021. С. 48.

12. Turova P., Stavrianidi A. Various machine learning methods in HPLC-MS datasets treatment // The 13th Winter symposium on Chemometrics. Abstract Book. Москва, Россия. 2022. С. 42.

Благодарности

Автор выражает признательность и благодарность научному руководителю к.х.н. Ставрианиди А.Н. за помощь в постановке задач и обсуждении результатов исследования; д.х.н., проф., член-корр. РАН Шпигуну О.А.; д.х.н. Родину И.А.; д.х.н. Гречникову А.А.; к.б.н. Лыскову Д.Ф.; Подольскому И.И. за помощь и консультации по тематике работы; всем членам лаборатории хроматографии и лаборатории масс-спектрометрии за помощь в работе и поддержку.