DOI: https://doi.org/10.26896/1028-6861-2020-86-7-12-19

АЛГОРИТМ СОЧЕТАНИЯ ХРОМАТО-МАСС-СПЕКТРОМЕТРИЧЕСКОГО НЕНАПРАВЛЕННОГО ПРОФИЛИРОВАНИЯ И МНОГОМЕРНОГО АНАЛИЗА ДЛЯ ВЫЯВЛЕНИЯ ВЕЩЕСТВ-МАРКЕРОВ В ОБРАЗЦАХ СЛОЖНОГО СОСТАВА

© **Иван Викторович Плющенко[1]*, Дмитрий Геннадьевич Шахматов[2], Игорь Александрович Родин[1]**

[1] Московский государственный университет имени М. В. Ломоносова, Химический факультет, Россия, 119991, Москва, ГСП-1, Ленинские горы, 1, стр. 3; *e-mail: plyushchenko.ivan@gmail.com

[2] Государственный научный центр колопроктологии им. А. Н. Рыжих Минздрава России, Россия, 123423, Москва, ул. Саляма Адиля, 2.

Статья поступила 13 апреля 2020 г. Поступила после доработки 13 апреля 2020 г. Принята к публикации 27 мая 2020 г.

Лавинообразное развитие методов статистической обработки данных, вычислительных мощностей, техники хромато-масс-спектрометрического анализа и омиксных технологий в последние десятилетия так и не привело к созданию унифицированного протокола для ненаправленного профилирования. Влияние систематических ошибок снижает воспроизводимость и достоверность результатов исследования, одновременно затрудняя объединение и анализ данных масштабных многодневных хромато-масс-спектрометрических экспериментов. В работе предложен алгоритм проведения омиксного профилирования для выявления потенциальных веществ-маркеров в образцах сложного состава на примере анализа образцов мочи разных клинических групп пациентов. Профилирование проведено методом жидкостной хромато-масс-спектрометрии. Выбор маркеров проводили методами многомерного анализа, в том числе машинного обучения и отбора переменных. Тестирование подхода выполняли с использованием независимого набора данных алгоритмами кластеризации и проецирования на главные компоненты.

Ключевые слова: жидкостная хроматография; масс-спектрометрия; метаболомика; многомерный анализ; хемометрика; машинное обучение.

ALGORITHM OF COMBINING CHROMATOGRAPHY MASS SPECTROMETRY-UNTARGETED PROFILING AND MULTIVARIATE ANALYSIS FOR IDENTIFICATION OF MARKER-SUBSTANCES IN SAMPLES OF COMPLEX COMPOSITION

© **Ivan V. Plyushchenko[1]*, Dmitry G. Shakhmatov[2], Igor A. Rodin[1]**

[1] M. V. Lomonosov Moscow State University, Chemistry Department, 1/3 Leninskiye Gory, Moscow, 119991, Russia; *e-mail: plyushchenko.ivan@gmail.com

[2] State Scientific Center of Coloproctology, 2 ul. Salyama Adilya, Moscow, 123423, Russia.

Received April 13, 2020. Revised April 13, 2020. Accepted May 27, 2020.

A viral development of statistical data processing, computing capabilities, chromatography-mass spectrometry, and omics technologies (technologies based on the achievements of genomics, transcriptomics, proteomics, metabolomics) in recent decades has not led to formation of a unified protocol for untargeted profiling. Systematic errors reduce the reproducibility and reliability of the obtained results, and at the same time hinder consolidation and analysis of data gained in large-scale multi-day experiments. We propose an algorithm for conducting omics profiling to identify potential markers in the samples of complex composition and present the case study of urine samples obtained from different clinical groups of patients. Profiling was carried out by the method of liquid chromatography mass spectrometry. The markers were selected using methods of multivariate analysis including machine learning and feature selection. Testing of the approach was performed using an independent dataset by clustering and projection on principal components.

Keywords: liquid chromatography; mass spectrometry; metabolomics; multivariate analysis; chemometrics; machine learning.

Введение

В течение трех последних десятилетий развиваются так называемые «омиксные» технологии. Наиболее впечатляющие результаты достигнуты в области системной биологии [1]. Описание живых организмов на трех последовательных уровнях (геномном, протеомном, метаболомном) позволяет наиболее полно изучить воздействие экспериментальных факторов изменчивости на биологическую систему. Омиксные подходы, основанные на хромато-масс-спектрометрическом ненаправленном профилировании [2], завоевывают все большую популярность в последнее десятилетие, и число новых приложений постоянно растет. В качестве наиболее распространенных стоит упомянуть: гербаломику [3], петраломику [4], фудомику [5], типизацию нефтепродуктов и топлив [6], контроль качества лекарственных средств [7].

Хромато-масс-спектрометрическое ненаправленное профилирование служит для поиска веществ-маркеров, позволяющих проводить классификацию образцов по экспериментальным группам методами статистического анализа. К главным этапам исследования относятся: постановка задачи и отбор образцов нескольких классов, подготовка проб, хромато-масс-спектрометрическое профилирование, интегрирование хроматограмм и отбор веществ-маркеров методами статистического анализа (чаще всего — дискриминантного). Единого протокола обработки и проведения ненаправленного омиксного профилирования до сих пор не разработано. Можно выделить две ключевые проблемы. Одна из них — порядок выполнения объединенного расчета результатов экспериментов из нескольких исследований и аналитических последовательностей [8]. Совместное действие систематических ошибок отбора пробы и анализа наряду с влиянием случайных факторов вызывает смещение в данных, а также снижает воспроизводимость и достоверность результатов. Для повышения интерпретируемости данных необходимо максимально нивелировать влияние нежелательных факторов. Следующее затруднение — коррекция сигнала масс-спектрометрического детектора [9]. Необходимость этой процедуры связана с неустранимым загрязнением источника ионизации и постоянно действующим искажением сигнала при анализе.

Целью настоящего исследования стала разработка подхода для выявления потенциальных веществ-маркеров. Предложенный алгоритм должен, по возможности, максимально эффективно удалять устранимые систематические смещения в данных и обеспечивать минимизацию искажения сигнала детектора. В качестве объекта исследования были проанализированы образцы мочи, отобранные у разных клинических групп пациентов. Профилирование образцов выполняли с использованием хромато-масс-спектрометрической системы. Отбор потенциальных веществ-маркеров проводили с применением методов многомерного анализа и машинного обучения. Применимость подхода контролировали с использованием независимого набора данных алгоритмами хемометрического моделирования.

Экспериментальная часть

Реактивы и оборудование. В работе использовали муравьиную кислоту (98 %, хч) и ацетонитрил (для ВЭЖХ) производства Panreac (Испания), а также деионизированную воду (Milli-Q, Millipore, США). В качестве основы для приготовления раствора внутреннего стандарта применяли раствор для инъекций «Папаверин Буфус» (Производственная фармацевтическая компания «Обновление», Россия).

Анализ проводили с использованием хроматографа 1290 Infinity II (Agilent Technologies, США) с автоматическим дозатором проб и термостатируемым отделением для проб. Температура автосэмплера составляла 4 °C. Детектором служил масс-спектрометр (МСД) типа «тройной квадруполь» 6470 (Agilent Technologies, США), оснащенный источником электрораспылительной ионизации (ИЭР) JetSpray (Agilent Technologies, США). Сбор данных и первичную обработку хроматограмм проводили с помощью программного обеспечения MassHunter B.08.02. (Agilent Technologies, США). Хроматографическое разделение осуществляли на колонке Aquity UPLC BEH C18 (Waters, США) на основе гидрофобизированного силикагеля с параметрами 2,1 мм × 100 мм и диаметром частиц сорбента 1,7 мкм. Для увеличения срока службы хроматографических колонок использовали универсальные предколонки для ВЭЖХ SecurityGuard C18 (Phenomenex, США).

Для центрифугирования образцов использовали центрифугу СМ-50 (Elmi, Латвия). Для конвертирования хроматограмм в формат .mzXML применяли программное обеспечение ProteoWizard [10].

Интегрирование пиков, их выравнивание по временам и массам в разных хроматограммах и составление общей таблицы пиков осуществляли с помощью программы iMet-Q [11].

Статистический анализ, обработку данных и машинное обучение проводили в программной среде R (версия 3.6.1) [12]. Для чтения и сохранения таблиц использовали функции пакета data.table. В целях ускорения проводили мультиядерные вычисления (5 виртуальных ядер) с

помощью пакетов parallel, doParallel. Для одномерного статистического анализа и заполнения пропущенных значений использовали базовые пакеты base, stats; для машинного обучения — пакет caret [13]; для отбора переменных — пакет vscc [14], отбора повторяющихся предикторов — tuple. Обучение без учителя проводили с использованием пакетов factoextra, FactoMineR, валидацию кластеризации — NbClust, cluster, rafalib. Для создания графики применяли пакеты ggsci, cowplot, reshape2, ggplot2. Программный код, информация о сессии и пакетах доступны по адресу: https://github.com/plyush1993/Multistudy-experiments-with-Multiple-batches.

Коррекцию сигнала МСД проводили на сервере NOREVA [15] и с помощью пакетов vsn [16], ProteoMM [17], affy [18].

Подготовка и хранение образцов. Образцы хранили в морозильной камере Liebherr G 1213 (Liebherr, Германия) при рабочей температуре –25 °С.

Все анализируемые образцы мочи были предоставлены ФГБУ «Государственный научный центр колопроктологии имени А. Н. Рыжих» Министерства здравоохранения Российской Федерации. Образцы поступали двумя партиями: первую партию отбирали осенью – зимой 2016 г., вторую — летом – осенью 2017 г. Таким образом, весь эксперимент был разделен на две части: первую партию проанализировали весной 2017 г. (эксперимент 1), вторую — осенью 2017 г. (эксперимент 2). Краткое описание партий приведено в табл. 1.

Первая партия содержала всего 40 образцов: 20 образцов были отобраны до проведения операции (у пациентов с подтвержденным диагнозом колоректального рака), 8 — у контрольной группы (КГ) (пациенты с неонкологическими заболеваниями желудочно-кишечного тракта) и 12 — после проведения операции. Вторая партия содержала 48 образцов: 22 были отобраны до операции, 26 — у контрольной группы. Для приготовления образца контроля качества (КК) смешивали в равных долях все образцы из партии, перемешивали и переносили в индивидуальные емкости, замораживая вместе с образцами. В день анализа использовали новый образец КК.

Таблица 1. Характеристики партий образцов
Table 1. Characteristics of the sample sets

Номер эксперимента	Группа	Количество образцов
1	До	20
	После	12
	КГ	8
2	До	22
	КГ	26

Перед анализом образцы размораживали при комнатной температуре, затем отбирали аликвоту и центрифугировали при 16 000 мин$^{-1}$ в течение 15 мин. Надосадочную жидкость отбирали, помещали в другую емкость, разбавляли деионизованной водой в пять раз (с добавкой раствора внутреннего стандарта — папаверин Буфус — до конечной концентрации 0,2 мкг/мл) и повторно центрифугировали при 16 000 мин$^{-1}$ в течение 15 мин.

Проведение анализа. Определение проводили с использованием источника ИЭР в режиме регистрации положительно заряженных ионов по полному ионному току при следующих условиях: диапазон масс — 100 – 850 *m/z*, время сканирования — 500 мс, напряжение фрагментора — 130 В, ускоряющее напряжение в ячейке соударений — 7 В, температура осушающего газа — 325 °С, поток осушающего газа — 10 л/мин, давление распыляющего газа — 45 psi, температура газа оболочки — 350 °С, поток газа оболочки — 12 л/мин, температура источника — 350 °С, напряжение капилляра — 3500 В, напряжение сопла — 500 В. Температура термостата колонки составляла 30 °С, объем вводимой пробы — 5 мкл. Разделение пробы проводили в градиентном режиме подачи элюента, скорость потока составляла 0,3 мл/мин. Программа бинарного градиентного элюирования: 0 мин 5 % В, 5 мин 5 % В, 35 мин 59 % В, 42 мин 95 % В, 52 мин 95 % В, 55 мин 5 % В, 60 мин 5 % В (фаза А — деионизованная вода с 0,1 % муравьиной кислоты, фаза В — ацетонитрил). При этом для уменьшения загрязнения камеры ионизации в первые 2,5 мин и последние 15 мин поток из колонки направляли на слив.

Первый ввод пробы служит для уравновешивания системы и проверки ее чистоты — это всегда был холостой образец (вода). Следующим анализируемым образцом был КК. Затем вводили 5 анализируемых образцов, затем воду, снова 5 образцов и после образец КК. Такой цикл повторяли необходимое число раз. Ввод холостого образца между анализами необходим для проверки чистоты системы, КК — для проверки ее стабильности. В конце последовательности еще раз анализировали образец контроля качества, затем хроматографическую систему промывали смесью ацетонитрил:вода (95:5) в течение 3 ч. После каждой последовательности иглу электроспрея и камеру ионизации очищали смесью изопропанол:вода (50:50) в соответствии с рекомендациями производителя. Калибровку масс-анализатора проводили перед каждой аналитической последовательностью. Эксперимент 1 был выполнен за 5

аналитических последовательностей, эксперимент 2 — за 8.

Для проверки стабильности системы сравнивали образцы КК между собой как внутри одной последовательности, так и между разными. В качестве первичной оценки служило наложение хроматографических профилей в режиме полного ионного тока. Затем сравнивали площади и времена удерживания пиков на хроматограммах по базовым ионам. Отклонение времен удерживания должно быть не более 30 с, а площадей — не более 10 %.

Для каждого образца проводили два последовательных определения. Пики на хроматограммах по базовым ионам должны максимально совпадать (абсолютное отклонение времен элюирования — менее 10 с, площадей пиков — менее 10 %).

Обсуждение результатов

Расчет таблицы пиков. Для интерпретации полученных хромато-масс-спектрометрических профилей была сгенерирована таблица пиков. Полученные хроматограммы конвертировали в формат .mzXML с помощью программного обеспечения ProteoWizard. В программе MSConvert были установлены фильтры по подбору пиков (PeakPicking) для 1-го уровня масс-спектров и маркер заголовка (TitleMarker).

Для интегрирования и разметки пиков использовали программное обеспечение «iMet-Q». Диапазон времен выравнивания пиков (rtTol) определяли по разнице времен элюирования папаверина в первом и последнем образце: он составил 0,4 мин. Разница во временах удерживания объясняется загрязнением колонки или модификацией поверхности сорбента. При этом разница во временах элюирования может не быть одинаковой для всех соединений. Для более надежного установления диапазона дрейфа времен удерживания сравнивали времена удерживания соединений на хроматограммах по базовым пикам в образцах КК, проанализированных в разных аналитических последовательностях. Максимальная разница во временах элюирования для всех пиков была не более 0,4 мин. Поэтому был выбран диапазон в 0,4 мин, соответствующий максимальному разбросу времен удерживания. Диапазон выравнивания по отношению масса/заряд (mzTol) определяется исходя из характеристик масс-спектрометра, в первую очередь — разрешения. Для современных МСД типа «тройной квадруполь» эта величина составляет около 0,1 Да. Параметр mzWidth определяется как ширина пика (в Да): этот параметр подбирали эмпирически, критерием служило сравнение интенсивностей пиков внутреннего стандарта (папаверина) во всех образцах в iMet-Q и MassHunter. Наилучшее совпадение наблюдалось при значении mzWidth, равном 0,1 Да. Соотношение сигнал/шум было установлено по умолчанию и составляло 3 ед. Диапазоны масс и времен элюирования соответствовали условиям анализа (100 – 850 Да, 2,5 – 45 мин). Результаты интегрирования для повторных измерений усредняли автоматически, параметры интегрирования и выравнивания пиков были одинаковыми для обоих экспериментов.

Выбор метода коррекции МСД. Эксперимент 2 использовали для тренировки и настройки моделей, а также отбора потенциальных веществ-маркеров, так как объем выборки и сбалансированность классов в нем выше, чем в эксперименте 1. Эксперимент 1 служил проверочным для оценки применимости подхода.

На первом этапе проводили заполнение пропущенных значений по алгоритму половины минимального значения по признаку для каждой таблицы пиков. Затем отбирали наилучшие методы коррекции сигнала в терминах минимального относительного стандартного отклонения (ОСО) внутри каждой аналитической последовательности и между всеми на сервере NOREVA.

Таблица 2. Результаты сравнения методов коррекции МСД в терминах ОСО
Table 2. Comparison of MSD correction procedures in terms of RSD

Тип коррекции МСД	Процент от общего числа пиков с ОСО, меньшим или равным указанного значения				
	10 %	15 %	20 %	30 %	50 %
«Сырые» данные	1	7	30	51	65
Auto Scaling	30	30	30	30	30
Cubic Spline	69	69	69	69	69
Cyclic Loess	20	32	42	54	66
EigenMS	69	69	69	69	69
MSTUS	5	21	35	51	65
Quantile	69	69	69	69	69
VSN	69	69	69	69	69

Для этого в таблице пиков эксперимента 2 были отобраны только образцы КК. Так как образцы КК во всех последовательностях были одинаковыми, число признаков (пиков) с минимальным ОСО должно быть максимальным. В табл. 2 приведены результаты сравнения всех последовательностей данных по проценту пиков от общего числа с ОСО, меньшим или равным 10, 15, 20, 30 и 50 %.

Анализируя данные табл. 2, можно заключить, что потенциально оптимальными методами коррекции сигнала стоит считать VSN, Quantile, Cubic Spline и EigenMS. Они обеспечивают ОСО, меньшее или равное 10 %, почти для 70 % всех пиков в образцах КК.

Следующий этап отбора метода коррекции МСД — оптимизация по точности классификации. Для правильной коррекции недостаточно снизить значения среднего ОСО, необходимо также убедиться, что полезная информация будет сохранена. Для этого проводили расчет для четырех выбранных методов коррекции в эксперименте 2 в среде R.

Затем для каждой из четырех таблиц с методом коррекции МСД проводили фильтрацию предикторов (пиков) по p-значению (ОСФ — одномерный статистический фильтр). Первый этап ОСФ — тест нормальности Шапиро – Уилка и тест гомогенности дисперсии Бартлетта. Если p-значение для предиктора в тесте Шапиро – Уилка меньше 0,05, вычисляли непараметрический тест Уилкоксона с поправкой Бенджамини – Хохберга на множественные сравнения и оставляли признаки с p-значением, меньшим 0,05. Аналогичным образом проводили тест Уэлча (в случае негомогенной дисперсии и нормального распределения) и тест Стьюдента (нормальное распределение, гомогенная дисперсия). Во всех случаях уровень фильтрации признаков был установлен по p-значению равным 0,05 с поправкой Бенджамини – Хохберга.

Следующий шаг — расчет точности классификации для моделей статистического обучения. Применяли 5-блочную кросс-валидацию с тремя повторами, оптимизацию проводили по показателю точности классификации. Были выбраны четыре модели машинного обучения и в каждой — один параметр для оптимизации в диапазоне пяти значений: к-ближайших соседей (KNN, параметр оптимизации — число соседей), проекция на латентные структуры (PLS, число компонент), машина опорных векторов с радиальной ядерной функцией (SVM, параметр C), случайный лес (RF, число предикторов для дерева). Таким образом, настройку каждой модели проводили 75 раз (пять блоков кросс-валидации с тремя повторами и еще пять раз для параметра оптимизации модели). В табл. 3 приведены показатели точности классификации моделей машинного обучения для четырех наборов данных после применения разных методов коррекции МСД и для «сырых» данных. Из данных табл. 3 следует, что оптимальным методом коррекции МСД является EigenMS (все методы демонстрируют точность выше 85 %).

Отбор веществ-маркеров. Для отбора веществ-маркеров сначала отсортировали 50 самых важных предикторов в каждой модели (для каждого алгоритма используется своя метрика определения важности переменных). Затем объединили все 200 признаков в один список и выделили наиболее стабильные предикторы, т.е. те, которые встречаются минимум дважды (СПФ), и провели автоматический отбор переменных (АОП) по алгоритму vscc. Алгоритм vscc проводит одновременную минимизацию внутригрупповой дисперсии при максимальной межгрупповой дисперсии. Сочетание этих двух критериев автоматически выделяет переменные, которые лучше всего разделяют образцы между группами. Параллельно с применением описанных алгоритмов на каждой стадии проводили независимый контроль процедуры сокращения признакового пространства. Для этого вычисляли среднее для всех переменных значение площади под ROC-кривой (AUROC): если на каждой стадии удаляли только неинформативные предикторы,

Таблица 3. Точность классификации (%) для наборов данных после проведения методов коррекции МСД с применением различных методов

Table 3. Accuracy of classification (%) for datasets after implementation of different MSD correction procedures

Метод коррекции	Модель машинного обучения			
	KNN	SVM	RF	PLS
Quantile	51	67	80	70
Cubic Spline	50	69	80	69
VSN	80	87	82	87
EigenMS	86	99	96	91
«Сырые» данные	67	71	81	98

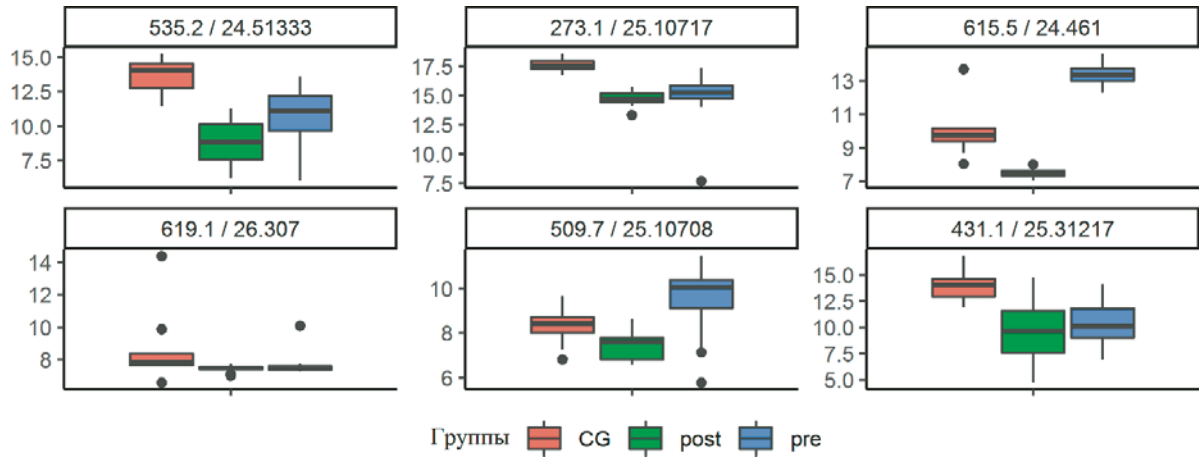

Рис. 1. Диаграммы размаха признаков из усеченного набора эксперимента 1: CG — контрольная группа; pre — до операции, post — после операции (обозначения предикторов: *mz*/время удерживания, мин)

Fig. 1. Box plots for features from a reduced dataset of experiment 1: CG — control group, pre — before operation, post — after operation. Predictor notation: *mz*/retention time, min

среднее значение AUROC увеличивалось после каждой стадии.

Валидация результатов. После отбора переменных в эксперименте 2 провели расчет коррекции EigenMS для таблицы пиков эксперимента 1. Затем вручную отобрали вещества-маркеры в эксперименте 1 из выбранных по алгоритму vscc в эксперименте 2. Выбор осуществляли по значению *m*/*z* (с точностью 0,1 Да) и временам удерживания (с точностью 0,3 мин). Из 11 веществ-маркеров, выбранных после проведения статистического анализа в эксперименте 2, удалось обнаружить 6 компонентов в таблице пиков эксперимента 1. Вероятно, обнаружить все вещества-маркеры не удалось из-за затруднений в интегрировании пиков для образцов сложного состава, а также из-за возможных различий в процедуре сбора образцов. Полученный усеченный набор данных использовали для дальнейшей валидации. Диаграммы размаха для каждого соединения в усеченном наборе эксперимента 1 приведены на рис. 1.

Объем выборки слишком мал для надежной оценки распределения признаков между группами, однако видно, что во всех случаях 25 и 75 процентили отличаются по крайней мере между двумя группами, а для двух маркеров (509.7/25.10708 и 615.5/24.461) — между всеми группами. Характеристики основных процедур многомерного статистического анализа, примененных последовательно для каждого эксперимента, приведены в табл. 4.

Из табл. 4 следует, что каждый этап статистической обработки последовательно уменьшает число предикторов при постоянном росте среднего значения AUROC и точности классификации. Эти обстоятельства подтверждают применимость используемых процедур для уменьшения признакового пространства и выделения наиболее информативных переменных.

Таблица 4. Характеристики основных этапов статистического анализа

Table 4. Characteristics of the main stages of statistical analysis

Набор данных		Число предикторов	Среднее AUROC	Точность классификации			
				KNN	SVM	RF	PLS
Эксперимент 2	«Сырые» данные	4174	0,55	67*	71*	81*	98*
	EigenMS	4174	0,558	86	99	96	91
	ОСФ	341	0,71				
	СПФ	51	0,851	85	97	98	96
	АОП	11	0,908	91	96	94	90
Эксперимент 1	«Сырые» данные	3440	0,683	60*	67*	95*	76*
	EigenMS	3440	0,802	100*	100*	100*	100*
	EigenMS + АОП	6	0,933	96	92	93	98

Примечание. Приведена точность классификации после применения ОСФ.

На финальном этапе проверки применимости разработанного подхода проводили многомерное проецирование методами обучения без учителя (рис. 2) и валидацию кластеризации (рис. 3).

График счетов первых двух главных компонент для усеченного набора данных эксперимента 1 (см. рис. 2, *а*) позволяет визуально оценить расположение образцов. Наблюдения каждой группы расположены близко друг к другу и пространственно отделены от других групп, и, хотя есть несколько выбросов, большую часть образцов можно верно классифицировать по графику. Дендрограмма наблюдений (см. рис. 2, *б*) позволяет однозначно и правильно классифицировать все образцы по трем группам. Разница высоты рассечения дендрограммы для двух и трех клас-

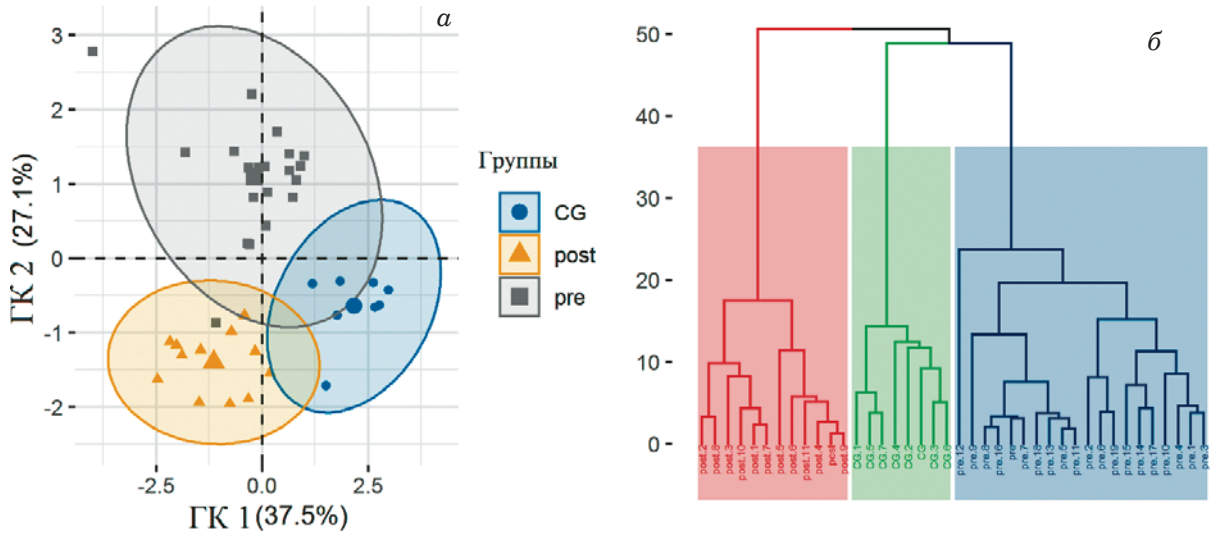

Рис. 2. Многомерное проецирование методами обучения без учителя для эксперимента 1: *а* — проекция наблюдений на главные компоненты (цвета соответствуют группам: CG — контрольная группа, pre — до операции, post — после операции; эллипсы соответствуют доверительной вероятности 95 %); *б* — дендрограмма наблюдений: расстояние манхэттена, объединение по алгоритму Уорда (цвета листьев соответствуют их классу в исходном наборе данных; названия листьев аналогичны наблюдениям исходного набора данных; ветви дендрограммы и области вокруг них окрашены по условию наличия трех групп в данных: кластеризация проведена для трех классов)

Fig. 2. Multidimensional projection via methods of unsupervised learning for experiment 1: *a* — projection of observations on the principal components (colors correspond to the groups: CG — control group, pre — before the operation, post — after the operation, ellipses — correspond to 95% confidence interval); *b* — dendrogram of observations (Manhattan distance, aggregation by Ward algorithm (ward.D2)). Colors of the leaves correspond to their class in the initial dataset; leaf names correspond to the observations of the original dataset; branches of the dendrogram and the areas around them are colored according to the condition of three groups present in the dataset (clustering was carried out for three classes)

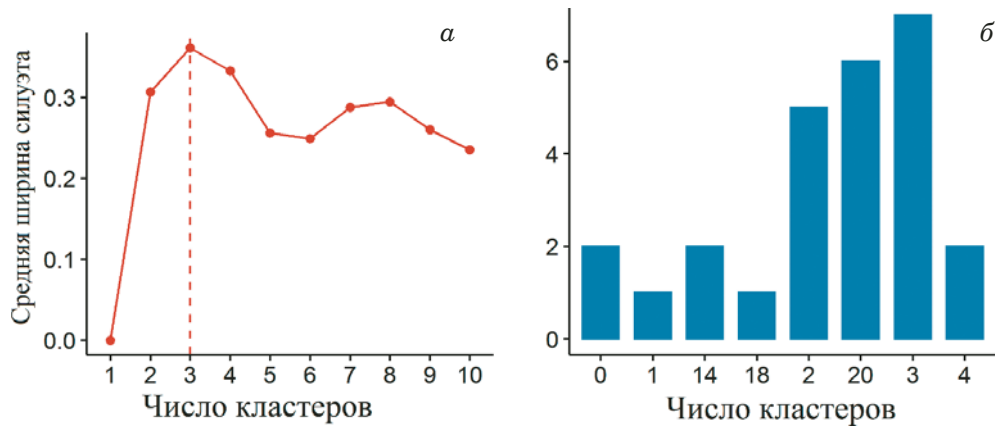

Рис. 3. Результаты валидации кластеризации по алгоритму среднего силуэта (*а*) и гистограмма распределения статистики результатов определения числа кластеров по 26 индексам (*б*) (расстояние манхэттена, объединение по алгоритму Уорда (ward.D2))

Fig. 3. The results of clustering validation according to the algorithm of an average silhouette (*a*) and histogram of the distribution of statistics on the results of determination of the number of clusters by 26 indices (*b*) (Manhattan distance, aggregation by Ward algorithm (ward.D2))

сов менее пяти единиц. Разница высоты рассечения дендрограммы для трех и четырех классов более 20 единиц. Дальнейшее разбиение на классы дает разницу менее 5 – 10 единиц. Даже без изначальной гипотезы о числе групп в наборе данных структура дендрограммы указывает, что наиболее предпочтительное число классов равно трем, как и в исходном наборе эксперимента 1. Дополнительно была проведена проверка кластеризации независимыми методами (см. рис. 3).

Из результатов расчета статистики среднего силуэта (см. рис. 3, *а*) следует, что оптимальное число кластеров равно трем, как и в наборе данных. Аналогичные результаты дает оценка числа кластеров по 26 различным метрикам (см. рис. 3, *б*, описание метрик см. в документации пакета NbClust, функция NbClust аргумент index = «all»).

Заключение

Таким образом, разработан и апробирован на независимом наборе данных подход для определения потенциальных веществ-маркеров, позволяющий проводить многомерную классификацию образцов сложного состава. Предложенный подход может быть воспроизводимо реализован и использован для любых исследований с хромато-масс-спектрометрическим определением для выборок разного объема. В качестве потенциальных приложений наиболее востребованными представляется анализ образцов и материалов многокомпонентного состава в целях определения характеристичных компонентов для систематизации и контроля качества.

Финансирование

Работа поддержана Российским фондом фундаментальных исследований (РФФИ) (№ гранта: Аспиранты 19-33-90071).

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

ЛИТЕРАТУРА (REFERENCES)

1. **Arivaradarajan P., Misra G.** (Eds.). Omics approaches, technologies and applications: integrative approaches for understanding OMICS data. 1st edition. — Singapore: Springer Nature, 2018. P. 158. DOI: 10.1007/978-981-13-2925-8_4.
2. **Gorrochategui E., Jaumot J., Lacorte S., Tauler R.** Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow / TrAC. 2016. Vol. 82. P. 425 – 442. DOI: 10.1016/j.trac.2016.07.004.
3. **Argueso C. T., Assmann S. M., Birnbaum K. D., et al.** Directions for research and training in plant omics: Big Questions and Big Data / Plant direct. 2019. Vol. 3. N 4. P. e00133. DOI: 10.1002/pld3.133.
4. **Lozano D. C. P., Thomas M. J., Jones H. E., Barrow M. P.** Petroleomics: Tools, Challenges, and Developments / Annu. Rev. Anal. Chem. 2020. Vol. 13. P. 20. 1 – 20. 26. DOI: 10.1146/annurev_anchem-091619-091824.
5. **Ferranti P.** The future of analytical chemistry in foodomics / Curr. Opin. Food Sci. 2018. Vol. 22. P. 102 – 108. DOI: 10.1016/j.cofs.2018.02.005.
6. **Bolotnik T. A., Timchenko Y. V., Plyushchenko I. V.** Use of Chemometric Methods of Data Analysis for the Identification and Typification of Petroleum and Petroleum Products / J. Anal. Chem. 2019. Vol. 74. N 13. P. 1336 – 1340. DOI: 10.1134/S1061934819130045.
7. **Kharyuk P., Nazarenko D., Oseledets I., et al.** Employing fingerprinting of medicinal plants by means of LC-MS and machine learning for species identification task / Sci. Rep. 2018. Vol. 8. N 1. P. 17053. DOI: 10.1038/s41598-018-35399-z.
8. **Cui X., Tang J., Yang Q., et al.** Assessing the effectiveness of direct data merging strategy in long-term and large-scale pharmacometabonomics / Front. Pharmacol. 2019. Vol. 10. P. 127. DOI: 10.3389/fphar.2019.00127.
9. **Yang Q., Hong J., Li Y., et al.** A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies / Brief. Bioinform. 2019. DOI: 10.1093/bib/bbz137.
10. **Holman J. D., Tabb D. L., Mallick P.** Employing ProteoWizard to convert raw mass spectrometry data / Curr. Protoc. Bioinformatics. 2014. Vol. 46. N 1. P. 13.24.1 – 13.24.9. DOI: 10.1002/0471250953.bi1324s46.
11. **Chang H. Y., Chen C. T., Lih T. M., et al.** iMet-Q: a user-friendly tool for label-free metabolomics quantitation using dynamic peak-width determination / PLOS one. 2016. Vol. 11. N 1. P. e0146112. DOI: 10.1371/journal.pone.0146112.
12. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2019.
13. **Kuhn M., Johnson K.** Applied predictive modeling. 1st edition. — New York: Springer, 2013. — 615 p. DOI: 10.1007/978-1-4614-6849-3.
14. **Andrews J. L., McNicholas P. D.** Variable selection for clustering and classification / J. Classif. 2014. Vol. 31. N 2. P. 136 – 153. DOI: 10.1007/s00357-013-9139-2.
15. **Li B., Tang J., Yang Q., et al.** NOREVA: normalization and evaluation of MS-based metabolomics data / Nucleic Acids Res. 2017. Vol. 45. N W1. P. W162 – W170. DOI: 10.1093/nar/gKx449.
16. **Huber W., Von Heydebreck A., Sültmann H., et al.** Variance stabilization applied to microarray data calibration and to the quantification of differential expression / Bioinformatics. 2002. Vol. 18. N 1. P. S96 – S104. DOI. 10. 1093/bioinformatics/18.suppl_1.S96.
17. **Karpievitch Y. V., Taverner T., Adkins J. N., et al.** Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition / Bioinformatics. 2009. Vol. 25. N 19. P. 2573 – 2580. DOI: 10.1093/bioinformatics/btp426.
18. **Gautier L., Cope L., Bolstad B. M., Irizarry R. A.** Affy-analysis of Affymetrix GeneChip data at the probe level / Bioinformatics. 2004. Vol. 20. N 3. P. 307 – 315. DOI: 10.1093/bioinformatics/btg405.