
**МАСС-СПЕКТРОМЕТРИЯ ДЛЯ БИОТЕХНОЛОГИИ.
ИНТЕРПРЕТАЦИЯ ДАННЫХ, МЕТОДОЛОГИЯ, ПРИМЕНЕНИЕ**

УДК 621.384.668.8: 577.112.6

© В. В. Макаров, Я. И. Лютвинский, А. Н. Веренчиков

**АЛГОРИТМ IPЕХ-2D ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ
О КОМПОНЕНТАХ ПРОБЫ ИЗ МАССИВОВ ДАННЫХ
(ВЭЖХ-МС)-ЭКСПЕРИМЕНТОВ ПРОТЕОМИКИ**

Описан алгоритм обработки хромато-масс-спектрометрических данных с целью извлечения информации о молекулярных массах, времени хроматографического элюирования и интенсивности сигналов компонентов пробы. Алгоритм адаптирован для обработки масс-спектров многокомпонентных смесей пептидов, характерных для экспериментов протеомтики. Описан исследовательский комплекс программ, в котором реализован данный алгоритм. Приводятся результаты тестирования алгоритма на масс-спектрометрических данных, полученных в модельном эксперименте по идентификации белка.

ВВЕДЕНИЕ

Особенностью масс-спектрометрических экспериментов протеомтики является высокая сложность анализируемых смесей (порядка 10^5 – 10^6 компонентов с содержанием в диапазоне от 10^{-15} до 10^{-4} М). Поэтому перед проведением масс-спектрометрического анализа проба, как правило, подвергается предварительному фракционированию и разделению методами двумерного гелеэлектрофореза и (или) высокоэффективной жидкостной хроматографии.

Перспективные разработки в области масс-спектрометрической техники нацелены на создание приборных комплексов с многоступенчатым разделением пробы, на повышение скорости разделения, разрешающей способности и чувствительности масс-спектрометрического анализа. Повышение этих характеристик сопряжено с ростом потоков экспериментальных данных, что в свою очередь требует повышения производительности методов первичной обработки и интерпретации масс-спектрометрических данных. Выполнение этих операций должно производиться за время, сопоставимое с продолжительностью самого эксперимента.

Ключевой задачей обработки масс-спектрометрических данных является извлечение аналитически значимой информации о компонентах пробы, на основе которой производится химико-биологическая интерпретация данных. Существующие алгоритмы либо встроены в коммерческие продукты и недоступны, либо доступны, но недостаточно эффективны, т. е. обладают низкой производительностью и плохо приспособлены для автоматической обработки масс-спектров многокомпонентных проб. Актуальной является разработка нового класса алгоритмов, обладающих высокой производительностью и позволяющих автоматизировать процесс обработки данных.

Важной инженерной задачей является создание программной среды для разработки и тестирования алгоритмов, в которой должны быть эффективно организованы хранение и доступ к масс-спектрометрическим данным, а также их визуализация.

В настоящей работе описан алгоритм IPЕХ-2D, предназначенный для обработки массива хромато-масс-спектрометрических данных с целью извлечения аналитически значимой информации о компонентах пробы. Приводится описание комплекса программ "Масс-процессор", в котором реализован данный алгоритм.

АЛГОРИТМ

Для решения задачи извлечения аналитически значимой информации из масс-спектра смеси пептидов был разработан алгоритм IPЕХ (сокращение от "Isotopic Pattern Extraction" — извлечение изотопных мультиплетов), подробно описанный в наших предыдущих работах [1–3]. Алгоритм IPЕХ-2D является модификацией алгоритма IPЕХ для массива данных (ВЭЖХ-МС)-эксперимента, задача обработки которого несколько отличается от обработки одиночного масс-спектра.

При (ВЭЖХ-МС)-эксперименте производится последовательная регистрация масс-спектров в процессе хроматографического элюирования компонентов пробы. Период регистрации Δt_{ch} выбирается с таким расчетом, чтобы за время выхода хроматографического пика компонента пробы было зарегистрировано как минимум 7–10 масс-спектров. Это позволяет фиксировать форму хроматографического пика данного компонента пробы с точностью, достаточной для определения положения центроида (центра тяжести пика).

Сигнал каждого из компонентов пробы распределен между несколькими последовательно зарегистрированными масс-спектрами, поэтому перед обработкой алгоритмом IPeX производится суммирование фрагментов сигнала, относящихся к каждому из компонентов пробы. Это позволяет повысить соотношение "сигнал/шум" и улучшить качество масс-спектрометрических данных.

Методы избирательного суммирования хромато-масс-спектрометрических данных разрабатывались в работах [4, 5] в приложении к приборным комплексам "газовый хроматограф—масс-спектрометр". Непосредственное применение данных методов для обработки массива (ВЭЖХ-МС)-данных затруднено необходимостью учета изотопных и зарядовых распределений ионов. Для обработки таких данных более эффективным является подход, предложенный в [6] для решения аналогичной задачи и основанный на "скользящем суммировании" массива масс-спектров.

На рис. 1 представлена схема обработки массива хромато-масс-спектрометрических данных. На начальном шаге производится расчет суммы S_Σ масс-спектров, зарегистрированных на промежутке времени, равном времени элюирования хроматографического пика $T_\Sigma = T_{ch}$. Время T_{ch} может быть рассчитано на основе характеристик используемого хроматографа и характеристик вещества, либо известно из предыдущих экспериментов. Далее осуществляется последовательная обработка массива масс-спектров S , на каждом шаге которой к сумме $S_\Sigma[i]$ добавляется масс-спектр, соот-

ветствующий времени элюирования $t_{ch} = T_\Sigma + \Delta t_{ch} \cdot (i + 1)$, и вычитается масс-спектр времени элюирования $t_{ch} = \Delta t_{ch} \cdot i$.

Каждый из суммарных масс-спектров $S_\Sigma[i]$ обрабатывается алгоритмом IPeX (шаг 2, рис. 1). Результатом обработки является список параметров компонентов пробы, элюированных в промежутке времени $(\Delta t_{ch} \cdot i; \Delta t_{ch} \cdot i + T_\Sigma)$. Данная информация включается в массив предварительных результатов обработки данных C^* .

Поскольку каждый из выделенных компонентов массива $C^*[i]$ может присутствовать в результатах обработки предшествующих и последующих суммарных масс-спектров, массив C^* содержит многократно дублированные результаты по каждому из компонентов пробы. Для устранения этого дублирования на массиве C^* решается задача кластерного анализа, в результате чего дублирующие друг друга элементы группируются в кластеры. Кластеры массива C^* имеют характер цепочек, элементы которых распределены по времени хроматографического выхода и сгруппированы по молекулярной массе. Для объединения компонентов в кластеры применяется невзвешенный центроидный метод, согласно которому компонент $C^*[i]$ с молекулярной массой $m_{C^*[i]}$ присоединяется к кластеру, расстояние до центра которого $|m_C - m_{C^*[i]}|$ имеет минимальную величину.

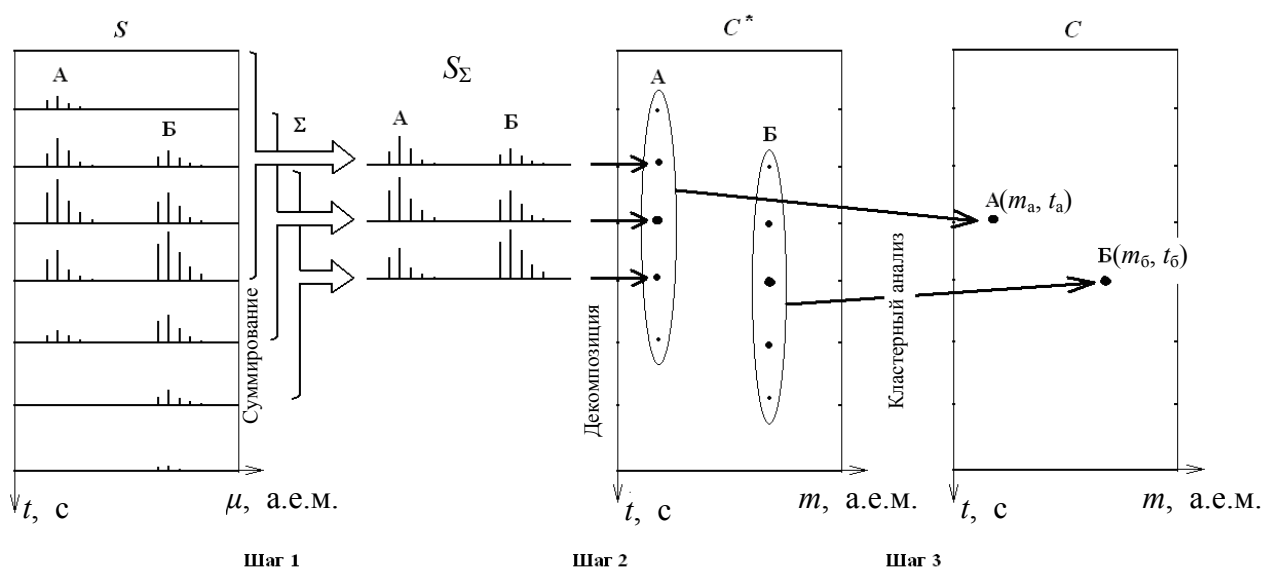


Рис. 1. Схема обработки массива хромато-масс-спектрометрических данных алгоритмом IPeX-2D

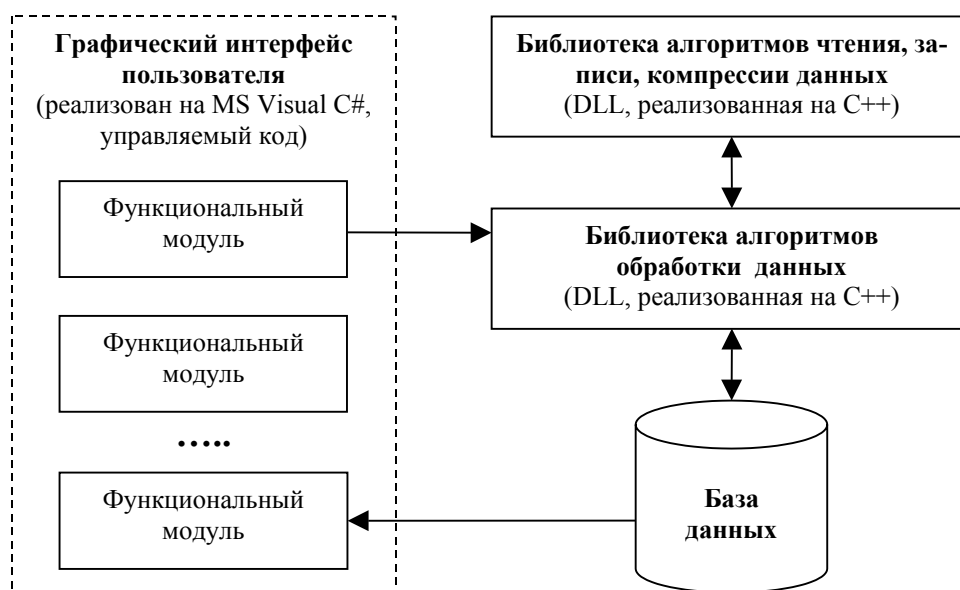


Рис. 2. Структурная схема комплекса программ

Присоединение компонента $C^*[i]$ к кластеру $C[j]$ производится в том случае, если разность соответствующих им молекулярных масс не выходит за пределы, определяемые разрешающей способностью масс-спектрометра R , а разность времени элюирования меньше установленного значения $T_{\max}$:

$$\begin{cases} |m_{C^*[i]} - m_{C[j]}| < \frac{m_{C^*[i]}}{R}, \\ |t_{C^*[i]} - t_{C[j]}| < T_{\max}. \end{cases} \quad (1)$$

После присоединения очередного компонента положение центра кластера (m_c, t_c) пересчитывается для учета параметров нового компонента:

$$(m_c)_n = \frac{m_c \cdot I_c + m_{C^*[i]} \cdot I_{C^*[i]}}{I_c + I_{C^*[i]}},$$

$$(t_c)_n = \frac{t_c \cdot I_c + t_{C^*[i]} \cdot I_{C^*[i]}}{I_c + I_{C^*[i]}}, \quad (I_c)_n = I_c + I_{C^*[i]},$$

где индекс n обозначает обновленные значения параметров кластера.

В случае, если условие (1) не выполняется ни для одного из существующих кластеров, результат $C^*[i]$ выделяется в отдельный кластер.

По завершении обработки кластеры тестируют-

ся на предмет достоверности. Кластер признается достоверным, если его элементы покрывают диапазон времени элюирования шириной не менее $T_{\Sigma}/2$, в противном случае кластер отбраковывается. Это позволяет исключить кластеры, состоящие из неправдоподобно малого количества элементов, подавляющее большинство которых являются ложноположительными результатами.

После выполнения этих операций массив C содержит только достоверные кластеры, соответствующие сигналам компонентов пробы. Молекулярная масса и среднее время элюирования каждого из компонентов пробы вычисляются как средневзвешенные значения на множестве элементов кластера:

$$m_{C[i]} = \frac{\sum_{C^*[j] \in C[i]} m_{C^*[j]} I_{C^*[j]}}{\sum_{C^*[j] \in C[i]} I_{C^*[j]}},$$

$$t_{C[i]} = \frac{\sum_{C^*[j] \in C[i]} t_{C^*[j]} I_{C^*[j]}}{\sum_{C^*[j] \in C[i]} I_{C^*[j]}}.$$

Суммарная интенсивность сигнала ионов компонента пробы определяется по элементу кластера с максимальной интенсивностью:

$$I_{C[i]} = \max_{C^*[j] \in C[i]} (I_{C^*[j]}).$$

РЕАЛИЗАЦИЯ АЛГОРИТМА

Для реализации и тестирования алгоритма IPX-2D и других алгоритмов обработки масс-спектрометрических данных был разработан исследовательский комплекс программ "Масс-процессор". Создание данного комплекса программ позволило унифицировать форматы хранения и процедуры доступа к хромато-масс-спектрометрическим данным, а также реализовать инструменты их визуализации.

В состав комплекса (рис. 2) входят следующие функциональные блоки.

1. Реляционная база данных формата Microsoft SQL Server для хранения хромато-масс-спектрометрических данных и результатов их обработки.

2. Программа "Масс-процессор", в которой реализован графический интерфейс пользователя, средства доступа к данным, инструменты обработки данных, отображения и сохранения результатов.

3. Библиотеки методов обработки хромато-

масс-спектрометрических данных.

4. Набор вспомогательных модулей для импортирования данных, полученных на оборудовании различных производителей.

Модульная структура графического интерфейса позволяет напрямую связать алгоритмы обработки данных с инструментами визуализации, обеспечить гибкость и расширяемость комплекса программ. Графический интерфейс программы "Масс-процессор" представлен на рис. 3. Доступ к базе экспериментальных данных организован в окне 1 в виде структуры. Визуализация масс-спектра осуществляется на одномерной диаграмме 2, в окне которой реализован стандартный набор инструментов масштабирования и выбора отображаемого диапазона шкалы. Для визуализации массива хромато-масс-спектрометрических данных разработан инструмент двумерной растровой диаграммы 3, которая строится в осях "Отношение массы иона к заряду—Время хроматографического элюирования", где логарифм интенсивности сигнала кодируется оттенками серого цвета различной плотности.

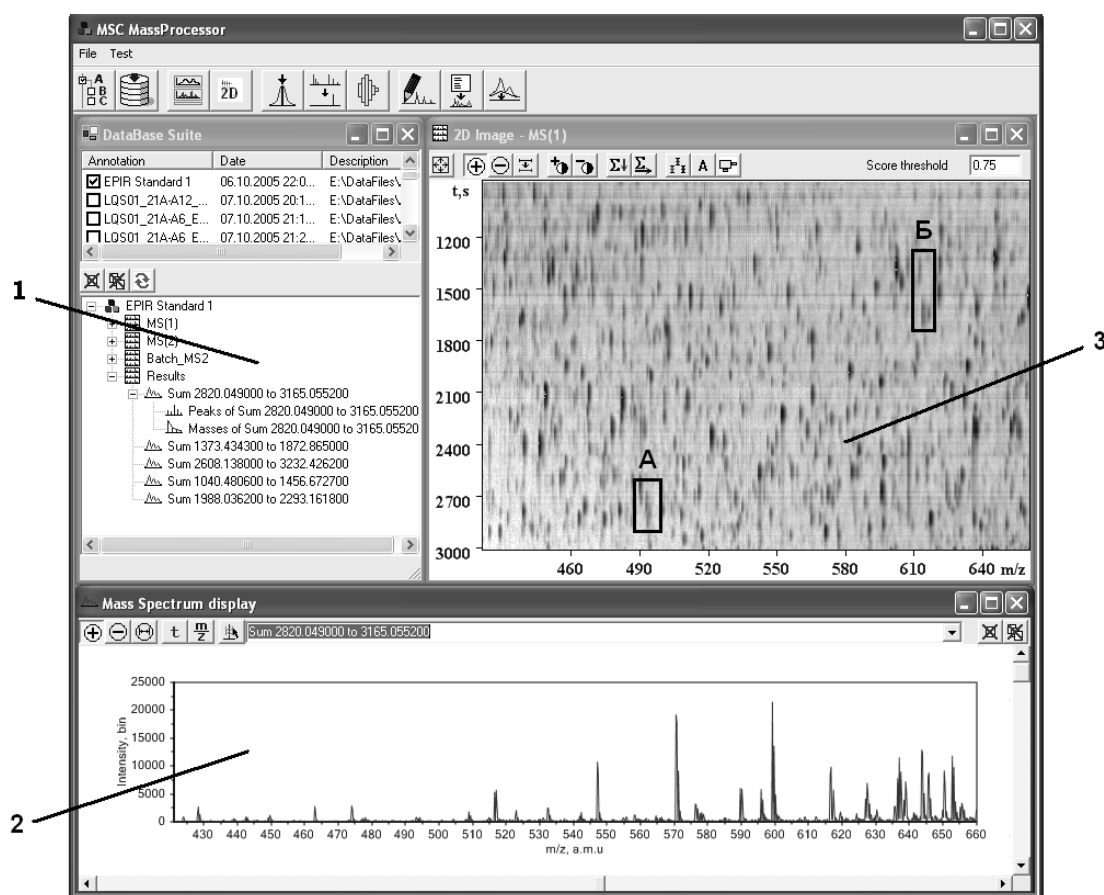


Рис. 3. Графический интерфейс программы "Масс-процессор". Пояснение в тексте

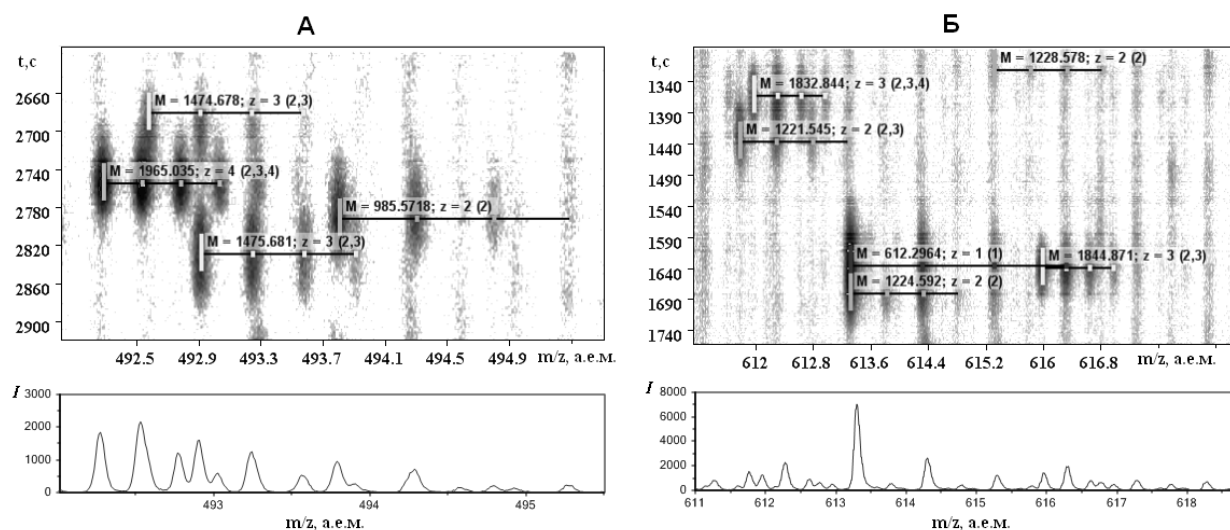


Рис. 4. Фрагменты массива хромато-масс-спектрометрических данных, представленного на рис. 3, и соответствующие суммарные масс-спектры

В графическом интерфейсе программы "Масс-процессор" на рис. 3 представлен массив хромато-масс-спектрометрических данных гидролизата модельной смеси белков (bovine serum albumin, bovin thyroglobulin, chicken ovalbumin, horse ferritin, fructose-bisphosphate aldolase, l-lactate dehydrogenase). На рис. 4 приведены результаты обработки данных алгоритмом IPHX-2D, иллюстрирующие способность алгоритма к выделению информации о компонентах пробы, сигналы которых претерпевают наложения в силу близких значений времени хроматографического элюирования и отношения массы к заряду. При визуальном анализе масс-спектров правильное выделение компонентов с молекулярными массами 1474.678, 1475.681, 1832.844, 1221.545 а.е.м. представляет значительные трудности по причине наложения спектральных пиков. При использовании алгоритма IPHX-2D данные компоненты были успешно выделены. Для каждого из перечисленных компонентов были обнаружены изотопные мультиплеты ионов нескольких зарядовых состояний, что свидетельствует о высокой достоверности выделения.

ИСПОЛЬЗОВАНИЕ АЛГОРИТМА IPHX-2D В ЭКСПЕРИМЕНТЕ ПО ИДЕНТИФИКАЦИИ БЕЛКА

Для тестирования алгоритма IPHX-2D был поставлен модельный эксперимент по идентификации белка методом пептидного массового картирования. В качестве модельного белка был выбран бычий сывороточный альбумин (БСА), широко

используемый для тестирования аналитических характеристик масс-спектрометрических приборов.

Препарат БСА (Sigma, США) был гидролизован трипсином, затем полученная смесь пептидов анализировалась методом ВЭЖХ-МС. Масс-спектрометрический анализ проводился на времяпролетном масс-спектрометре MX-5303 с источником ионов "электроспрей" (разработка Института аналитического приборостроения РАН). Предварительное разделение пробы производилось на жидкостном хроматографе "Милихром А-02" (производства ЗАО Институт хроматографии "Эконова", г. Новосибирск), работающем в режиме прямой стыковки с источником ионов.

В результате масс-спектрометрического эксперимента был получен массив из 569 масс-спектров, регистрация которых производилась в течение 21 мин с периодом около 2 с.

Массив экспериментальных данных был обработан алгоритмом IPHX-2D. Список из 367 компонентов пробы, полученный в результате обработки, был направлен в программу Mascot, доступную в сети Интернет по адресу <http://www.matrix-science.com> для интерпретации по методу пептидного массового картирования. Проба была верно интерпретирована как гидролизат БСА. Данный вариант интерпретации оценен наивысшим значением рейтинга, в то время как оценки прочих гипотез лежат ниже порога достоверности, вычисляемого программой Mascot.

Пептиды, обнаруженные в пробе, в совокупности покрывают 71 % аминокислотной последова-

тельности белка, что является высоким показателем для метода пептидного массового картирования. Высокая достоверность идентификации белка свидетельствует о качественном решении задачи извлечения аналитически значимой информации из масс-спектрометрических данных.

ЗАКЛЮЧЕНИЕ

Представленный алгоритм позволяет автоматизировать обработку хромато-масс-спектрометрических данных экспериментов протеомики. Высокая производительность алгоритма позволяет использовать его в программном обеспечении масс-спектрометрических приборов, разрабатываемых в настоящее время. Способность алгоритма к выделению информации о компонентах пробы, сигналы которых претерпевают наложения, позволяет успешно анализировать смеси с большим количеством компонентов, что особенно актуально для задач протеомики.

Авторы выражают благодарность Александру Подтележникову, Екатерине Подольской и Александру Новикову за предоставленные данные (ВЭЖХ-МС)-экспериментов.

СПИСОК ЛИТЕРАТУРЫ

1. Макаров В.В., Лютвинский Я.И., Савельев С.К., Веренчиков А.Н., Краснов Н.В. Алгоритм извлечения аналитически значимой ин-

формации из масс-спектрометрических данных экспериментов протеомики // Научное приборостроение. 2006. Т. 16, № 2. С. 92–100.

2. Makarov V.V., Saveliev S.K. Data mining method for electrospray mass spectra of peptide mixtures // Proceedings of SPIE. 2005. V. 6251. P. 355–362.
3. Макаров В.В., Самокиш А.В., Лютвинский Я.И. Метод извлечения значимой информации из масс-спектров пептидов // Научное приборостроение. 2004. Т. 14, № 2. С. 96–104.
4. Biller J.E., Biemann K. Reconstructed mass Spectra: A novel approach for the utilization of gas chromatograph-mass spectrometer data // Analytical Letters. 1974. V. 7, N 7. P. 515–528.
5. Dromey R.G., Stefik M.J., Rindfleisch T.C., Duffield A.M. Extraction of mass spectra free of background and neighboring component contributions from gas chromatography/mass spectrometry data // Analytical Chemistry. 1976. V. 48, N 9. P. 1368–1372.
6. Pearcy J.O., Lee T.D. MoWeD, a computer program to rapidly deconvolute low resolution electrospray liquid chromatography/mass spectrometry runs to determine component molecular weights // J. Am. Soc. Mass Spectrom. 2001. V.12, N 5. P. 599–606.

Институт аналитического приборостроения РАН, Санкт-Петербург

Материал поступил в редакцию 30.05.2006.

IPEX-2D: DATA MINING ALGORITHM FOR MASS SPECTRA OF LC-MS EXPERIMENT

V. V. Makarov, Ya. I. Lutvinsky, A. N. Verentchikov

Institute for Analytical Instrumentation RAS, Saint-Petersburg

An algorithm is described for extraction of analytically significant information from spectra of LC-MS experiment. The algorithm is adjusted for mass spectra of complex peptide mixtures, which are typical for proteomics experiments. The algorithm has been implemented in the research-grade software and tested in the example protein identification.