

УДК 621.384.668.8: 577.112.6

© В. Макаров, А. Самокиш, Я. И. Лютвинский

МЕТОД ИЗВЛЕЧЕНИЯ ЗНАЧИМОЙ ИНФОРМАЦИИ ИЗ МАСС-СПЕКТРОВ ПЕПТИДОВ

Описан метод обработки масс-спектров пептидов с целью извлечения информации о массах молекул пробы. Предметом обработки является массив данных о молекулярных пиках, выделенных на стадии предварительной обработки. Метод основан на анализе и декомпозиции структуры масс-спектра и предполагает использование статистических данных о форме изотопных распределений молекулярных пиков пептидов, которые получены методом численного моделирования на множестве идеализированных масс-спектров пептидов. Приводится описание алгоритма, реализующего предложенный метод, а также результаты обработки набора тестовых данных.

ВВЕДЕНИЕ

Результаты современного масс-спектрометрического эксперимента характеризуются большими потоками данных (порядка 10 спектров в 1 секунду) и требуют обработки в режиме реального времени. В задачах идентификации протеинов целью обработки масс-спектра является выделение масс пептидов, содержащихся в пробе, для последующего поиска протеинов в базах данных. Решение данной задачи производится путем декомпозиции масс-спектра, сложность структуры которого обусловлена изотопным распределением химических элементов пробы, особенностями процесса ионизации (наличием или отсутствием многозарядных ионов), присутствием в спектре химических и аппаратных шумов.

Высокая частота регистрации спектров предъявляет жесткие требования к времени выполнения процедур обработки. Предлагается высокоскоростной алгоритм обработки масс-спектра с целью извлечения информации о массах молекул пробы, основанный на структурной декомпозиции масс-спектра и использовании статистических данных.

1. ОБЩАЯ СХЕМА ОБРАБОТКИ РЕЗУЛЬТАТОВ МАСС-СПЕКТРОМЕТРИЧЕСКОГО ЭКСПЕРИМЕНТА

Схема обработки результатов масс-спектрометрического эксперимента в значительной степени определяется конфигурацией используемого аналитического оборудования и особенностями его работы. При использовании аппаратной конфигурации "жидкостный хроматограф—время-пролетный масс-спектрометр с ионным источником электроспрей" обработка результатов масс-спектрометрического анализа проводится в несколько этапов (рис. 1, 2).

Схема обработки результатов эксперимента



Рис. 1. Схема обработки результатов LS-MS-эксперимента. Предварительная обработка масс-спектра

В процессе эксперимента производится накопление спектральных данных, которые затем подвергаются первичной обработке (рис. 1). В масс-спектрах производится выделение молекулярных пиков, вычисляются положение их центроидов и относительная интенсивность. Полученный массив пиков характеризуется сложной структурой и представляет собой композицию зарядно-изотопных распределений молекулярных пиков химических соединений пробы. Алгоритм декомпозиции осуществляет поиск в масс-спектре структур, соответствующих изотопным распределениям молекулярных пиков, после чего вычисляются моноизотопные массы соответствующих молекул.

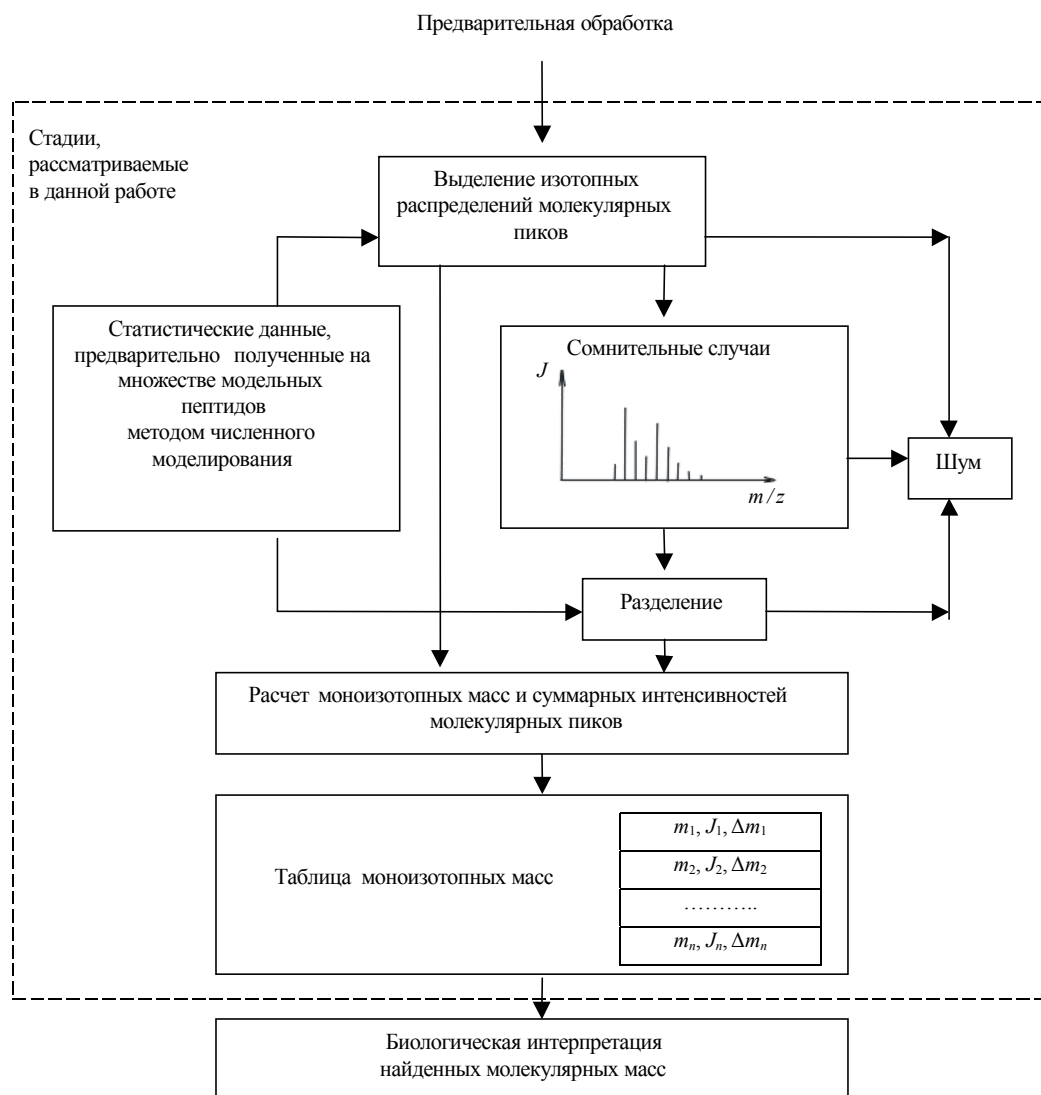


Рис. 2. Схема обработки результатов эксперимента.
Структурная обработка масс-спектра

В задачах идентификации протеинов предмет масс-спектрометрического анализа являются пептиды, представляющие собой цепочки аминокислот с молекулярной массой 300–6000 а.е.м., терминированные атомом водорода и гидроксигруппой. На рис. 3 представлены типичные изотопные распределения молекулярных пиков среднестатистического пептида в разных диапазонах масс. Моноизотопным пиком (МП, рис. 3) называется пик, образованный ионами, в которых все атомы представлены наиболее распространенным

изотопом. Идентификация данного пика позволяет определить молекулярную массу пептида. В изотопных распределениях пептидов моноизотопным является первый пик, однако в случае больших молекулярных масс (рис. 3, в) он обладает низкой интенсивностью, часто сравнимой с уровнем шума, и может быть потерян на стадии первичной обработки. Кроме того, встречаются ситуации, когда изотопные распределения нескольких ионов накладываются друг на друга (рис. 3, г, д, е). Подобные наложения затрудняют

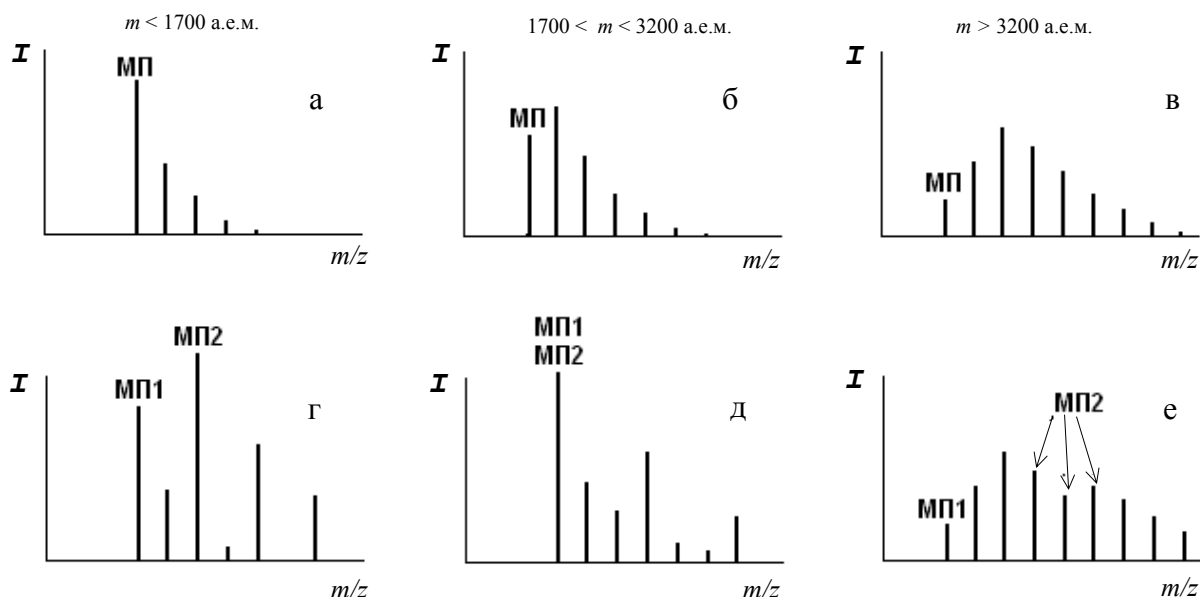


Рис. 3. Изотопные распределения молекулярного пика среднестатистического пептида (а, б, в) и группы пиков, наблюдаемые в экспериментальных спектрах: г — наложение изотопных распределений ионов с различными зарядами; д — то же, с наложением моноизотопных пиков МП1 и МП2; е — наложение изотопных распределений ионов с одинаковыми зарядами (выделение моноизотопного пика МП2 затруднено)

выделение моноизотопных пиков, а следовательно, и расчет молекулярных масс. Возникает необходимость детектирования групп пиков, обладающих вышеперечисленными особенностями, с целью предотвращения некорректных результатов декомпозиции.

Предлагается метод сравнения с эталоном, позволяющий произвести расчет моноизотопной массы молекулы независимо от особенностей группы молекулярных пиков, а также оценить соответствие группы пиков изотопному распределению молекулы определенного класса биологических соединений. В качестве дополнительных критериев соответствия предлагается использовать второй и третий центральный моменты группы пиков.

2. СТАТИСТИЧЕСКИЕ ПАРАМЕТРЫ ИЗОТОПНЫХ РАСПРЕДЕЛЕНИЙ МОЛЕКУЛЯРНЫХ ПИКОВ ПЕПТИДОВ

2.1. Критерии принадлежности группы пиков к изотопному распределению молекулы

Изотопное распределение молекулярного пика можно рассматривать как распределение случай-

ной величины "масса молекулы". Известно, что распределение случайной величины однозначно характеризуется совокупностью своих центральных моментов. Таким образом, каждой точке шкалы масс соответствует ряд статистических распределений центральных моментов изотопного распределения молекулярного пика. Из всей совокупности центральных моментов в данном случае интерес представляют второй центральный момент (дисперсия), характеризующий протяженность группы пиков, а также третий центральный момент изотопного распределения молекулярного пика, характеризующий его асимметрию:

$$K_1 = Dx = \sum_i (m_i - \mu)^2 w_i,$$

$$K_2 = M_3 = \sum_i (m_i - \mu)^3 w_i,$$

где w_i — относительная интенсивность i -го изотопного пика, m_i — масса пика, μ — средняя масса группы пиков.

В случае, если интенсивности пиков тестовой группы искажены наложением посторонних пиков, значения центральных моментов будут существенно отличаться от центров распределений

на данной массе (например, точка *Б* на рис. 4), выходя за пределы доверительного интервала. Таким образом, данные статистические характеристики служат критериями принадлежности группы пиков к изотопному распределению молекулы определенной массы.

2.2. Эталонная форма изотопного распределения молекулы пептида

Изотопное распределение молекулярного пика среднестатистического пептида обладает характерной формой (рис. 3, а, б, в), которая может быть описана векторами масс и относительных интенсивностей пиков:

$$\begin{aligned}\overline{M} &= \{m_1, m_2, \dots, m_n\}, \\ \overline{W} &= \{w(m_1), w(m_2), \dots, w(m_n)\},\end{aligned}$$

где

$$w(m_i) = \begin{cases} \frac{J(m_i)}{\sum_{k=1}^n J(m_k)}, & m_i \in \overline{M}; \\ 0, & m_i \notin \overline{M}, \end{cases}$$

$J(m_i)$ — интенсивность i -го изотопного пика.

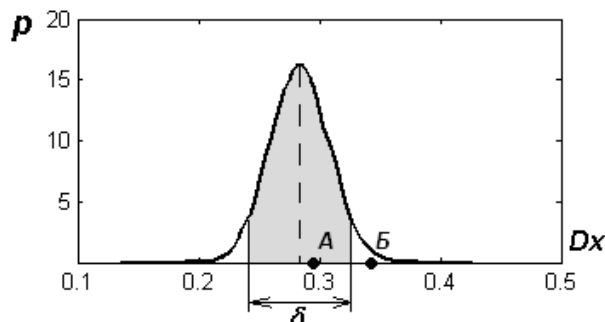


Рис. 4. Плотность распределения дисперсии изотопного распределения молекулы пептида с массой 2000 а.е.м.; δ — доверительный интервал

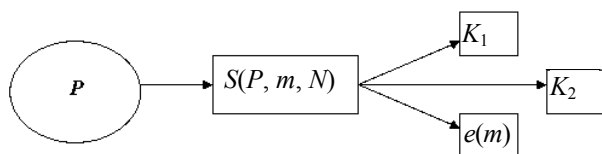


Рис. 5. Схема расчета статистических распределений критериев

Форма изотопного распределения молекулярного пика пептида определяется молекулярной массой и в меньшей степени составом молекулы. Каждой молекулярной массе m соответствует форма изотопного распределения среднестатистического пептида, которая может служить своеобразным эталоном. Сравнивая форму тестируемой группы пиков с эталоном, можно оценить гипотезу о том, что данная группа пиков является изотопным распределением пептида массы m^* . Критерием сравнения является минимальная сумма квадратов отклонений относительных интенсивностей пиков группы и эталона:

$$K_3 = \min(\kappa(p)), \quad (1)$$

где $\kappa(p) = \sum_{i=p}^{p+n} (w(m_i) - e(m_i))^2$; p — индекс позиции моноизотопного пика эталона, $p = 1, 2, \dots, N$; N — количество пиков масс-спектра; $e(m_i)$ — относительная интенсивность эталонного i -го изотопного пика.

Значению K_3 соответствует оптимальное значение индекса p , при котором наблюдается наилучшее совпадение эталона с тестируемой группой пиков. Тогда наиболее вероятным значением моноизотопной массы тестируемой группы пиков будет являться масса m_p первого изотопного пика эталона:

$$m^* = m_p. \quad (2)$$

Таким образом, моноизотопная масса тестируемой группы пиков определяется по первому (моноизотопному) пику эталона.

2.3. Расчет статистических параметров изотопных распределений молекулярных пиков пептидов

Статистические параметры, используемые для оценки гипотез о принадлежности пиков к изотопному распределению пептидов, получены численным расчетом, который производился методом Монте-Карло в несколько этапов (рис. 5). На первом этапе моделировалось множество P брутто-формул пептидов в диапазоне масс 300–6000 а.е.м. Пептиды, полученные в результате трипсинолиза, упрощенно представлены в виде цепочек аминокислот, заканчивающихся лизином или аргинином, терминированных с обоих концов атомом водорода и гидроксогруппой.

Затем для полученного множества P пептидов производился расчет множества S идеализированных изотопных распределений молекулярных пиков. На заключительном этапе на множестве S вычислялись следующие расчетные характеристики:

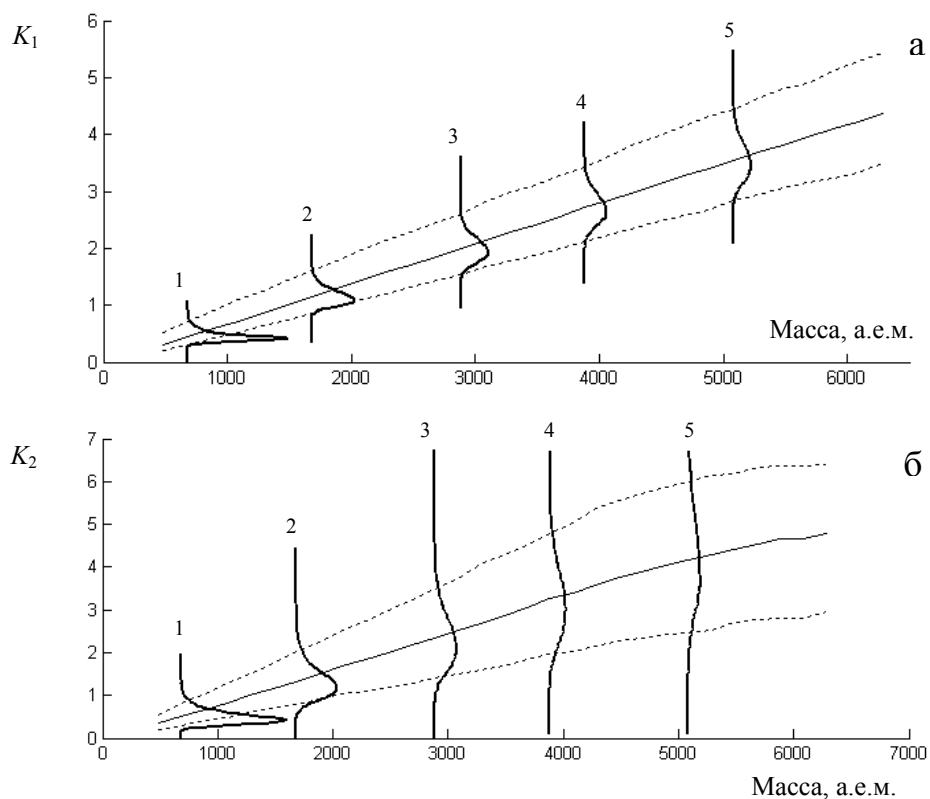


Рис. 6. Зависимость математического ожидания критериев от массы молекулы пептида на спектрах, образованных 500 ионами (а — критерий дисперсии изотопного распределения молекулярного пика, б — критерий третьего центрального момента изотопного распределения молекулярного пика). Пунктиром показаны границы доверительного интервала с вероятностью 0.95. 1, 2, 3, 4, 5 — плотности распределения критериев для ряда значений масс

— эталонные формы $e(m)$ изотопных распределений пептидов различных масс;

— зависимости критериев плотности распределения дисперсии $K_1(m)$ и третьего центрального момента $K_2(m)$ изотопных распределений пептидов от молекулярной массы.

На рис. 6 представлены зависимости математического ожидания исследуемых критериев от массы молекулы пептида, полученные численным расчетом. Вычисление производилось на массиве спектров, каждый из которых получен накоплением L ионов, где L — случайная величина, равномерно распределенная в диапазоне 500 ± 50 .

3. АЛГОРИТМ СТРУКТУРНОЙ ДЕКОМПОЗИЦИИ МАСС-СПЕКТРА

Каждое из содержащихся в пробе веществ представлено в масс-спектре несколькими группами пиков, в совокупности образующими зарядно-изотопное распределение ионов данного вещества. Структурная декомпозиция заключается в группировке пиков масс-спектра по принадлежности к ионам различных веществ.

Предлагаемый нами алгоритм включает две основные стадии. На первой стадии производится последовательный поиск изотопных распределений ионов различных зарядовых состояний. На за-

ключительной стадии производится группировка изотопных распределений, отнесенных к каждой из выявленных молекулярных масс. Совокупность найденных молекулярных масс описывает состав пробы.

3.1. Поиск изотопных распределений молекул пептидов в масс-спектре

Массив молекулярных пиков, выделенных на стадии предварительной обработки, сканируется на предмет поиска тестовых групп, которые могут быть интерпретированы как изотопные распределения молекул пептидов. Выделение тестовой группы происходит при выполнении ряда условий:

$$\left\{ \begin{aligned} \left(\mu_0 + \frac{k}{z} - \frac{\delta_i}{2} \right) \leq \mu_i < \left(\mu_0 + \frac{k}{z} + \frac{\delta_i}{2} \right); \\ \prod_{i=1}^{k_{\max}} w(\mu_i) \neq 0, \end{aligned} \right.$$

где $k = 0, \dots, (k_{\max} - 1)$, $k_{\max}$ — количество пиков тестовой группы, $\mu_0 = \frac{m_0}{z}$ — отношение массы к заряду для опорного пика, z — предполагаемый заряд иона, δ_i — погрешность определения центра пика i -го пика, вычисленная на стадии предварительной обработки.

В выделенной тестовой группе положения пиков на шкале масс приводятся в соответствие однозарядному иону, при этом используется допущение о том, что заряд иона обусловлен присоединением z протонов

$$m_{\text{тест}} = \mu_i \cdot z - m_{\text{H}} \cdot z,$$

где $k = 1, \dots, k_{\max}$, m_{H} — масса протона.

Оценка гипотезы о том, что тестовая группа является изотопным распределением молекулы пептида, производится на основе критериев, описанных в разделе 2. В случае если вычисленные значения критериев K_1 и K_2 попадают в пределы доверительного интервала, тестовая группа интерпретируется как изотопное распределение иона с зарядом z , моноизотопная масса которого вычисляется по формуле (2).

Если тестовая группа представляет собой наложение нескольких изотопных распределений, она обрабатывается процедурой разделения (см. пример на рис. 7). В этом случае позиционирование эталона $e_1(m)$ производится в соответствии с условием (1), причем сумма квадратов отклонений вычисляется только по первым n пикам тестовой группы, где n — наименьший порядковый номер пика, для которого выполняется условие

$$w(\mu_n) < w(\mu_{n-1}).$$

Выделенное таким образом изотопное распределение вычитается из тестовой группы. Пики с остаточными интенсивностями, превышающими пороговое значение $w_2(m) > \tau$, возвращаются в исходный массив. Затем производится поиск и интерпретация следующей тестовой группы (группа $w_2(m)$ на рис. 7). Описанный цикл обработки повторяется для всех зарядов состояний ионов в последовательности от больших зарядов к меньшим.

3.2. Результаты предварительного тестирования алгоритма декомпозиции

Вышеописанный алгоритм реализован в виде программного модуля системы обработки результатов эксперимента. Тестирование проводилось на массиве модельных и экспериментальных масс-спектров.

На рис. 8, 9 представлены результаты обработки экспериментального масс-спектра инсулина (*inculine porcine*, $m = 5773.6270$ а.е.м.), содержащего примесь грамицидина (*gramicidin-s*, $m = 1140.72148$ а.е.м.). В масс-спектре доминируют пики трех-, четырех- и пятизарядных ионов инсулина (см. рис. 8, фрагменты 2, 3, 4). Присутствие многочисленных модификаций инсулина с близкими молекулярными массами осложняет задачу декомпозиции, т. к. их группы пиков располагаются вблизи изотопного распределения ионов чистого

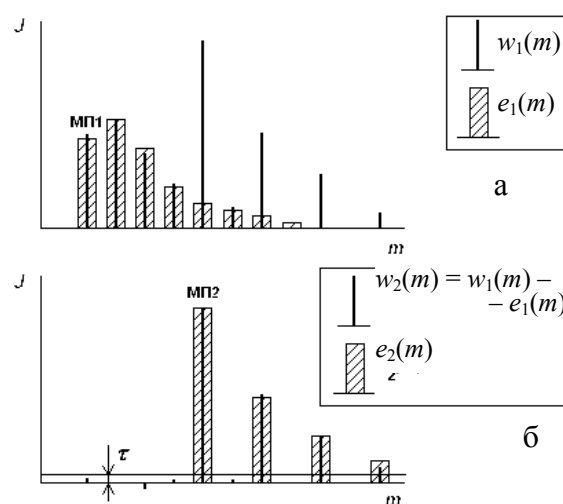


Рис. 7. Обработка тестовой группы, содержащей два наложенных друг на друга изотопных распределения

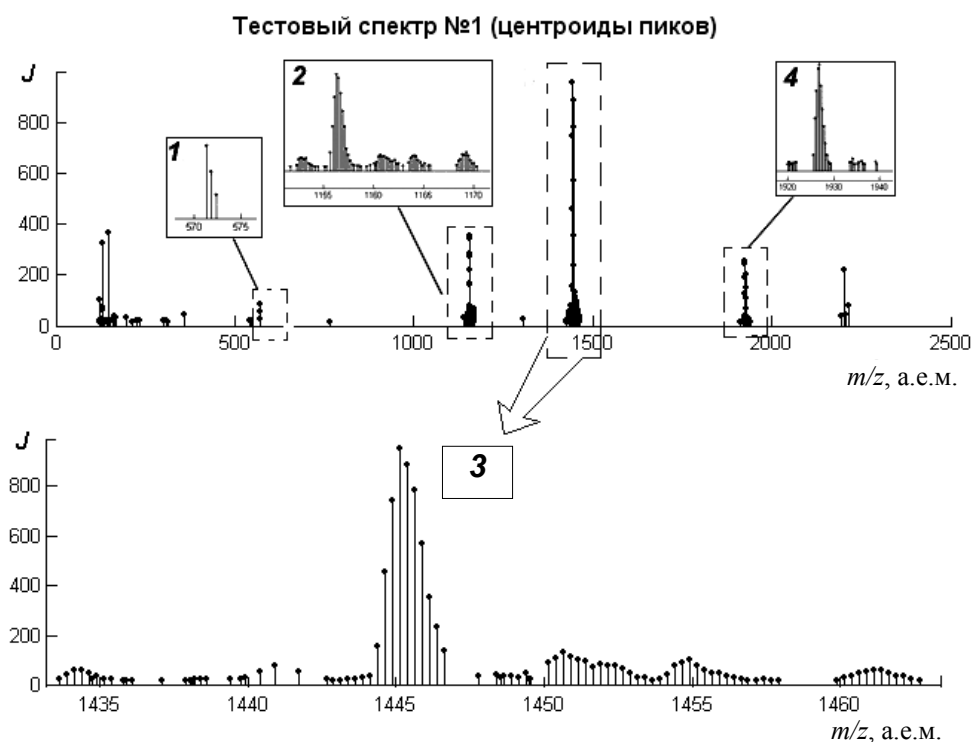


Рис. 8. Экспериментальный масс-спектр инсулина с примесью грамицидина.

Фрагменты спектра:

- 1 — изотопное распределение двухзарядного иона грамицидина;
- 2 — изотопные распределения пятизарядных ионов инсулина;
- 3 — изотопные распределения четырехзарядных ионов инсулина;
- 4 — изотопные распределения трехзарядных ионов инсулина



Рис. 9. Результат декомпозиции экспериментального масс-спектра инсулина с примесью грамицидина. Выделенные изотопные распределения молекул

инсулина и наслаиваются друг на друга. Грамицидин представлен одно- и двухзарядными ионами, пики которых имеют значительно меньшую интенсивность (см. рис.8, фрагмент 1). Обработка данного масс-спектра с применением вышеописанного алгоритма позволила определить массы основных компонент пробы. Помимо молекулярных масс инсулина и грамицидина были выявлены массы нескольких основных модификаций данных веществ (рис. 9), изотопные распределения которых наслаиваются друг на друга. В ходе декомпозиции было отсеяно большинство пиков химических и аппаратных шумов, содержащихся в масс-спектре. Время обработки данного спектра (246 выделенных пиков) составило 0.72 миллисекунды (на компьютере с процессором Intel Pentium 4, 2.4 ГГц).

На рис. 10, 11 представлен результат обработки масс-спектра смеси пептидов:

dioxycycline, $m = 444.1610$ а.е.м.;
gramicidin-s, $m = 1140.72148$ а.е.м.;
melittin, $m = 2844.75$ а.е.м.;
substance-p, $m = 1346.71$ а.е.м.;
insuline porcine, $m = 5773.6270$ а.е.м.

— компоненты представлены в равной концентрации.

В результате обработки были выявлены молекулярные массы всех компонент смеси, а также их модификаций. Особенностью данного спектра является избыток химических шумов, что осложняет задачу декомпозиции. Например, изотопное распределение двухзарядного иона массы $m = 1329.71$ а.е.м. (*substance-p* с потерей гидроксигруппы) сопровождается пиками шумов, сопоставимых по интенсивности с сигналом (рис. 11, фрагмент 2).

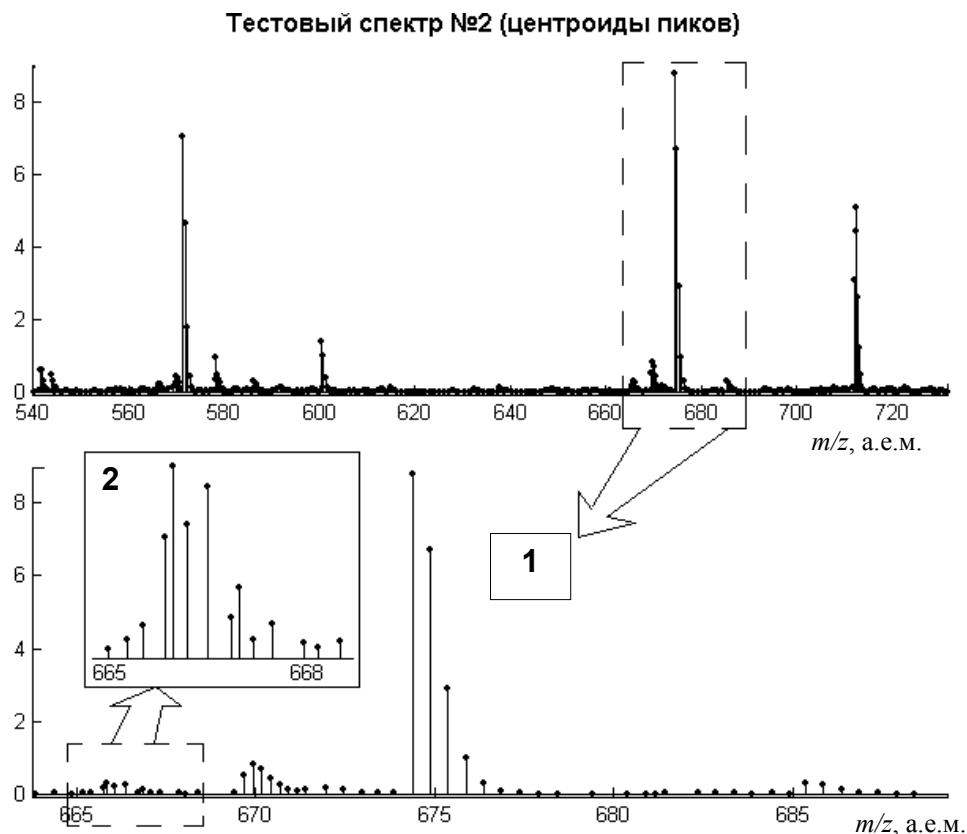


Рис. 10. Экспериментальный масс-спектр смеси пептидов. 1, 2 — вложенные фрагменты спектра. В фрагменте 2 — группа пиков, содержащая изотопное распределение двухзарядного иона массы $m = 1329.71$ а.е.м. (*substance-p* с потерей гидроксигруппы)

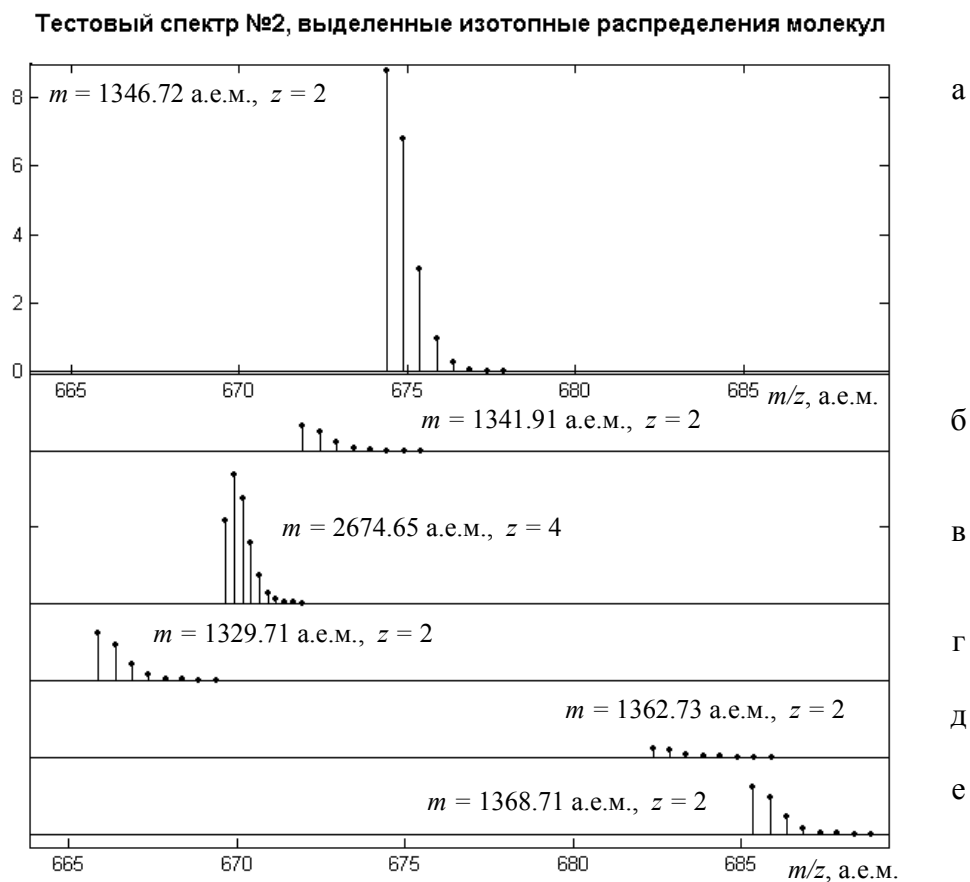


Рис. 11. Результат декомпозиции экспериментального масс-спектра смеси пептидов. Выделенные изотопные распределения молекул

В результате обработки данная масса была успешно выявлена (рис. 11, г) и посторонние пики были отсеяны, что свидетельствует о высокой селективности алгоритма. Время обработки данного спектра (1876 выделенных пиков) составило 14.21 миллисекунды (на компьютере с процессором Intel Pentium 4, 2.4 ГГц).

Результаты тестирования показали высокую эффективность предложенного алгоритма деком-

позиции, что свидетельствует о возможности его применения в программном обеспечении обработки результатов масс-спектрометрического эксперимента.

Институт аналитического приборостроения РАН, Санкт-Петербург

Материал поступил в редакцию 7.04.2004.

DATA MINING METHOD FOR PEPTIDE MASS SPECTRA

V. Makarov, A. Samokish, Y. I. Lutvinskiy

Institute for Analytical Instrumentation RAS, Saint-Petersburg

A data mining method performing isotope and charge state decomposition of peptide mass spectra is described. The method uses the peak table as an input and is based on the analysis and decomposition of the mass spectrum structure. Isotope cluster statistics is used as a base of decomposition. This statistic is calculated on a set of simulated mass spectra. The described method is programmed as a fast routine. The results of real spectra decomposition are presented.