УДК 543.51:621.384.8

МАТЕМАТИЧЕСКАЯ ОБРАБОТКА МАСС-СПЕКТРА С НЕ ПОЛНОСТЬЮ РАЗРЕШЕННЫМИ ПИКАМИ

А.В. Томилов, Б.А. Калинин, О.Е. Александров, В.Д. Селезнёв

Уральский государственный технический университет – УПИ имени первого Президента России Б.Н. Ельцина
Россия, 620002, Екатеринбург, ул. Мира, 19
tomilov@fizteh.org

Поступила в редакцию 17 ноября 2008 г.

Разработана программа, реализующая для магнито-секторного масс-спектрометра численные алгоритмы обработки масс-спектра, содержащего одиночные и не полностью разрешённые пары пиков. Рассмотрен алгоритм обработки.

Ключевые слова: магнитно-секторный анализатор, не полностью разрешённые пики, разрешение по массе, программная интерпретация масс-спектров, модель формы пика, аппроксимация формы пика

Томилов Анатолий Вячеславович - аспирант 1-го года обучения кафедры молекулярной физики физико-технического факультета УГТУ-УПИ, инженер кафедры информационных систем и технологий ИОИТ ГОУ ВПО УГТУ-УПИ.
Область научных интересов: масс-спектрометрия, системы реального времени, численные методы, приёмы программирования, формальные языки.

Калинин Борис Алексеевич - доцент кафедры молекулярной физики физико-технического факультета УГТУ-УПИ, кандидат физико-математических наук.
Область научных интересов: масс-спектрометрия.
Автор 30 статей.

Александров Олег Евгеньевич - доцент кафедры молекулярной физики физико-технического факультета УГТУ-УПИ, кандидат физико-математических наук.
Область научных интересов: масс-спектрометрия и теория разделения изотопов.
Автор 50 статей.

Селезнёв Владимир Дмитриевич - профессор кафедры молекулярной физики физико-технического факультета УГТУ-УПИ, доктор физико-математических наук.
Область научных интересов: динамика разреженных газов и неравновесная термодинамика.
Автор 200 научных работ.

Результатом масс-спектрометрического эксперимента является большой массив данных – зависимость интенсивности сигнала детектора от отношения массы иона к его заряду, с математической точки зрения определяющий функцию, заданную таблично. Задачей обработки данных является извлечение численными методами из этой функциональной зависимости аналитически значимой информации.

На масс-спектре визуально и при численной обработке можно выделить пики, соответствующие различным ионам. Ширина обособленного пика характеризует способность прибора различать близко расположенные пики и ограничивает возможности метода по анализу ионов с близкими значениями отношений массы к зарядовому числу. Наложение пиков вследствие их ненулевой ширины затрудняет определение их основных параметров и требует применение специальных методов математической обработки. Существует необходимость разработки программного обеспечения для реализации методов анализа в применении к масс-спектрам, содержащим не полностью разрешённые пики. Это связано с тем, что существующие программные продукты предлагают низкую степень автоматизации для решения задач этого класса. Специализированное же программное обеспечение производителей масс-спектрометрической техники ориентировано на аппаратную составляющую их продуктов.

Разработана программа bf, позволяющая осуществлять обработку обособленных пиков и пар не полностью разрешённых пиков, содержащихся в масс-спектрах. Данные для обработки должны быть представлены в виде файла-таблицы: пары значений отношений масс к зарядовым числам ионов $x = m/z$ и интенсивности сигнала $y = I$ детектора через символ-разделитель записаны построчно ANSI-текстом в порядке возрастания значений x:

$$\{((m/z)_j, I_j)\},\ j \in \{1, 2, \ldots, N\},$$

где j – номер строки, N – количество отсчётов.

Для извлечения аналитически значимой информации при анализе масс-спектров, содержащих не

полностью разрешённые пики, для различных методов разрешения (например, основанных на свёртке с производными функции [1, 2], описывающей форму пика, а также квазисплайновой деконволюции [3]) теряется информация о спектре на какой-либо стадии обработки. Метод свёрток требует предварительное сглаживание сигнала, например, вейвлет-фильтрацию [2, 4], которая подразумевает изменение оператором или каким-либо субъективным методом вейвлет-коэффициентов для обратного преобразования. Метод квази-сплайновой деконволюции существенно требует подбора основных параметров оператором, из-за чего теряется объективность метода. В связи с указанными недостатками для разрешения пиков было решено использовать метод без предварительной обработки анализируемого участка масс-спектра

Для нахождения параметров пиков принимается следующая модель спектра: спектр состоит из M пиков, каждому из которых соответствует слагаемое под знаком суммы в формуле

$$y(x) = n(x) + \sum_{i=1}^{M} a_i \cdot exp(-(f_r \cdot r_i \cdot \frac{x - m_i}{m_i})^2),$$

где $y = I$, $x = m/z$, $f_r = 2 \cdot \sqrt{ln(2)}$.

Слагаемое $n(x)$ заключает вклад фонового масс-спектра, свойства которого отличаются от свойств основного масс-спектра (пики существенно более плохо разрешаются прибором). Функция Гаусса в виде

$$f_i(x) = a_i \cdot exp(-(f_r \cdot r_i \cdot \frac{x - m_i}{m_i})^2), i \in \{0,1,...,M\}$$

описывает каждый отдельный пик таким образом, что все параметры имеют физический смысл и, кроме того, такой вид функции Гаусса удобен для использования в численных методах обработки. Параметр a_i – амплитуда пика, характеризует долю ионов в потоке. Параметр m_i – медиана спектра пика, за вычетом фона, соответствует m/z иона. Параметр r_i – величина, обратная частному ширины пика в единицах $dim\{m/z\}$ на половине высоты и значения величины m_i, характеризующая разрешающую способность масс-спектрометра. Число f_r – такой множитель для r_i, что значение r_i имеет указанный выше смысл. Постоянная f_r является положительным корнем уравнения

$$a_i \cdot exp(-(f_r \cdot r_i \cdot \frac{\Delta x_i/2}{m_i})^2) = \frac{a_i}{2},$$

где $\Delta x_i/2$ – половина ширины пика на половине его высоты. Паре пиков, таким образом, соответствует функциональная зависимость

$$y^*(x) = n^*(x) + a_1 \cdot exp(-(f_r \cdot r_1 \cdot \frac{x - m_1}{m_1})^2) + a_2 \cdot exp(-(f_r \cdot r_2 \cdot \frac{x - m_2}{m_2})^2).$$

Для нахождения параметров пиков на участке спектра алгоритм программы численными методами находит такие значения неизвестных параметров модели, при которых аппроксимация функциональной зависимостью рассматриваемого участка спектра считается наилучшей в соответствие с методом наибольшего правдоподобия.

Оценки, получаемые в методе наибольшего правдоподобия, обладают следующими важными свойствами:

1. Оценки состоятельны, то есть оценка неизвестного параметра стремится к истинному его значению при увеличении количества измерений.
2. Оценки асимптотически нормальны, то есть разброс оценки относительно истинного значения стремится к нормальному распределению при увеличении количества измерений.
3. Оценки асимптотически эффективны, то есть при любом другом способе обработки данных всегда имеет место больший разброс.
4. Оценки достаточны, то есть они используют максимум информации, содержащейся в обрабатываемых данных эксперимента.

Будем рассматривать работу секторного масс-спектрометра в режиме счёта ионов. Ионная статистика в этом режиме имеет некоторые особенности, при учёте которых можно получить из масс-спектра более полную информацию об анализируемом веществе.

Будем считать, что факт регистрации в детекторе каждого иона, достоверно вышедшего из источника ионов, является случайной физической величиной: исход каждого события регистрации независим от другого и зависит от факторов, которые мы не можем учесть. Тогда вероятность p для любого иона быть зарегистрированным постоянна, а исход этого события подчинён закону распределения Бернулли. Значение интенсивности для каждого отсчёта тогда есть случайная величина, так как само является суммой некоторого количества n случайных величин (фактов регистрации ионов) с распределением Бернулли, и, следовательно, распределено по биномиальному закону $Bin(n, p)$. При небольших ионных токах биномиальный закон будет соответствовать распределению Пуассона $Bin(n, p) \to P(n \cdot p)$, а при больших, согласно центральной предельной теореме, – распределению Гаусса $Bin(n, p) \to N(n \cdot p, n \cdot p \cdot (1 - p))$. Таким образом, для обычных спектров, снятых в режиме счёта, ошибки измерений интенсивности для каждого отсчёта распределены по нормальному закону. Характер ионной статистики наглядно подтверждается при рассмотрении масс-спектров, снятых приборами совершенно различных конструкций – на графиках можно визуально определить, что участки с большими интенсивностями сигнала имеют большее и абсолютное и относительное значение «дрожания» графика. Для отсчётов, снятых в различные моменты времени, в общем случае, не сохраняется ни абсолютная, ни относительная ошибки ($\sigma^2 = n \cdot p \cdot (1 - p) \Rightarrow \sigma \sim \sqrt{n}$). Измерения являются неравноточными.

В соответствие с принципом Лежандра, при нахождении наилучшей аппроксимации неравноточных экспериментальных данных, ошибки измерений которых распределены нормально, функция правдоподобия будет иметь максимум при значениях параметров аппроксимирующей функции, соответствующих минимуму суммы взвешенных квадратов невязок. Вес точки в серии измерений не равной точности с нормально распределёнными ошибками обратно пропорционален квадрату средней квадратичной ошибки соответственного измерения [5]. Средняя квадратичная ошибка серии равноточных измерений, под которыми подразумеваются регистрации ионов, пропорциональна корню квадратному из количества измерений в серии [5]. Таким образом, вес каждого отсчёта интенсивности

обратно пропорционален количеству сосчитанных ионов $p_j \sim 1/y_j$. Казалось бы, такой вес делает более значимыми точки спектра, которые лежат в областях между пиками и не несут информации об ионах, Это связано с видом функции, моделирующей форму пика.

Величина y_j значения интенсивности в режиме счёта, имеющая дискретный характер, может принимать нулевые значения для некоторых отсчётов. В этом случае, значения весов для таких точек необходимо задавать равными единице. Веса точек, определённые как функции от значений экспериментальных данных $p_j = 1/y_j$, содержат соответственные ошибки. Более точно веса точек можно определять, если брать значения интенсивностей более близкие к истинным значениям. Нетрудно показать, что в среднем по всем точкам таковыми являются значения $y_i(x_j)$, где вид функции y_i соответствует минимуму целевой функции. В таком виде в выражении для веса p_j точки исключается возможность ситуации, когда в ходе вычислений в знаменателе появляется нуль, ввиду положительной определённости функции y_i.

Численные эксперименты в ходе разработки программы показали, что учёт весов значительно улучшает практическую сходимость метода поиска при анализе пары пиков, один из которых по амплитуде существенно меньше другого. Это вызвано тем, что точки пика малой интенсивности имеют больший вес. Таким образом, учёт весов точек позволил повысить чувствительность метода.

Выбор пика (парного или одиночного) осуществляется оператором посредством выделения диапазона значений абсцисс $x = m/z$ на графике масс-спектра, который, по мнению оператора, полностью заключает интересующий пик. После этого во внимание принимаются только те точки, которые попали в эту область. Пусть выделен один одиночный пик под номером i: его спектр состоит из множества N точек $\{(x_j, y_j)\}, j \in \{1, 2, ..., N\}$. Будем считать, что спектр на этом участке приближённо описывается функцией

$$y_i(x) = n_i(x) + a_i \cdot exp(-(f_r \cdot r_i \cdot \frac{x - m_i}{m_i})^2),$$

при некоторых конкретных значениях параметров a_i, r_i, m_i и виде функции $n_i(x)$. Зависимость $n_i(x)$ принимается равной константе $n_i(x) = n_i = const$. Учитывая свойства фонового масс-спектра, считаем, что приближение константой является удовлетворительным для рассматриваемых узких участков спектра. При определении параметров пиков участка спектра мерой качества аппроксимации является целевая функция – сумма взвешенных квадратов невязок $\sum_{j=1}^{N} p_j \cdot (y_j - y_i(x_j))^2$, зависящая от вида функции y_i, где p_j – вес точки с номером j. Под зависимостью от вида функции y_i понимается зависимость от её параметров-констант n_i, r_i, m_i, a_i. Процедура формирующих пик. Однако на значения параметров пиков эти точки мало влияют, а влияют на значение параметров, описывающих фоновый масс-спектр $n(x)$.

нахождения минимума $\min_{y_i}(\sum_{j=1}^{N} p_j \cdot (y_j - y_i(x_j))^2)$ целевой функции может осуществляться любым численным методом поиска. В данной программе используется метод Хука-Дживса. Преимуществом этого метода является сильная практическая сходимость и простота программной реализации. Метод Хука-Дживса при анализе текущей точки в пространстве параметров поиска использует значения целевой функции в ней, но не её частных производных, нахождение и удовлетворительно точное вычисление значений которых численными способами затруднительно. Также в алгоритме метода не фигурируют вычисления детерминанта матрицы нормальных уравнений и её миноров, которые могут в общем случае становиться сингулярными в очередной итерации любого метода, использующего их текущие значения (например, метод Левенберга-Маккардта, используемый в TableCurve 2D).

Для начала работы метода поиска минимума целевой функции необходимо задать начальные значения параметров поиска. В нашем случае это n для участка спектра (n_i для участка спектра одного пика с номером i) и m_i, a_i, r_i для каждого пика, содержащегося в нём. Для расчёта первого приближения параметра n_i массив значений интенсивности $(y_1, y_2, ..., y_j, ..., y_N)$ упорядочивается по возрастанию значений y_j и вычисляется среднее арифметическое количества $[0.05 \cdot N]$ (целая часть числа) первых элементов полученного упорядоченного множества. Для упорядочивания используется алгоритм быстрой сортировки. В случае если N таково, что $[0.05 \cdot N] = 0$, то берётся только один первый элемент упорядоченного множества. Значение 0.05 выбрано произвольно. Оператор программы должен учитывать величину этого значения, когда выделяет одиночный пик, стараясь выделить достаточно широкую область. Для первого приближения m_i используется линейная интерполяция

$$x = x_p + (x_p - x_{p-1}) \cdot \frac{S_{p,N} - S_{1,N}/2}{S_{p-1,p}},$$

где p – наибольший номер точки такой, что $S_{p,N} > S_{1,N}/2$. В данном случае линейная интерполяция для m_i удовлетворительно точна. $S_{a,b}$ – площадь под графиком ломаной линии определяемой точками $(x_j, y_j), j \in \{1, 2, ..., N\}$ и осью абсцисс, заключённая между линиями $x = x_a, x = x_b$ и вычисленная по формуле трапеций

$$S_{a,b} = 0.5 \cdot (y_a \cdot (x_{a+1} - x_a) + \sum_{j=a+1}^{b-1} y_j \cdot (x_{j+1} - x_{j-1}) + y_b \cdot (x_b - x_{b-1})) \cdot$$

Первое приближение амплитуды вычисляется процедурой $a_i = max(y_1, y_2, ..., y_N)$. Первое приближение параметра r_i находится из уравнения

$$S_{a,b} \approx n_i \cdot (x_b - x_a) + \int_{-\infty}^{+\infty} a_i \cdot exp(-(f_r \cdot r_i \cdot \frac{x - m_i}{m_i})^2) dx ,$$

решением которого является

$$r_i \approx \frac{\sqrt{\pi / ln(2)}}{2} \cdot \frac{a_i \cdot m_i}{S_{a,b} - n_i \cdot (x_b - x_a)} .$$

Конечно, о состоятельности и несмещённости, а уж, тем более, об эффективности приведённых выше оценок, говорить не приходится (за исключением оценки m_i). Но, тем не менее, при тестировании в ходе разработки программы была определена полная пригодность таких оценок для использования.

Используя полученные оценки для параметров-констант функции y_i, производится дальнейшее уточнение их значений с использованием алгоритма минимизации целевой функции методом Хука-Дживса. Полученные в результате действий алгоритма оценки обладают всеми перечисленными выше свойствами оценок, получаемых в соответствие с принципом наибольшего правдоподобия.

В разработанной программе используется модификация метода Хука-Дживса с ограничениями, поэтому для корректной работы метода необходимо предварительно задать диапазоны изменения параметров-констант. Кроме того, необходимо задать начальные значения таких параметров метода поиска, как шаг поиска, его делитель и предельное значение шага поиска для каждого параметра-константы. Так как метод не чувствителен к значениям этих параметров, то их значения выбираются довольно грубо. Например, для амплитуды a_i ограничение снизу принимается равным нулю, а сверху – удвоенному значению начального значения амплитуды, начальное значение шага принимается равным одной сотой, а предельное значение – одной миллионной. Для шагов поиска всех переменных задаётся общий делитель равным тридцати. Введение ограничений позволяет сохранять свойство положительной определённости функции y_i в ходе процедуры поиска, а также гарантирует то, что параметры поиска не примут значений, не имеющих физического смысла.

Для анализа пары не полностью разрешённых пиков необходимо задание начальных значений семи параметров формулы для $y*$. Если пики разрешены настолько, что оператор может выделить их зоны, то в качестве начальных значений параметров для формулы двойных пиков используются значения, получаемые при обработке выделенных зон как одиночных пиков, иначе начальные значения задаёт оператор. Причём начальное значение параметра n вычисляется как среднее арифметическое значений соответственных параметров для зон левого и правого пиков. Далее к целевой функции для парного пика применяется метод минимизации. Полученные в результате его действий значения параметров и являются ответом к задаче анализа пары не полностью разрешённых пиков на участке масс-спектра.

Для оценки точности найденных значений параметров пиков подсчитываются наиболее вероятные значения средних квадратичных отклонений для всех параметров-переменных. Предварительно для этого вычисляются суммы взвешенных квадратов невязок и веса переменных. При определении весов переменных используются значения детерминантов матрицы нормальных уравнений и её миноров. При численных экспериментах в ходе разработки программы было установлено, что практическая вероятность того, что эти матрицы являются сингулярными для выбираемых участков спектров на конечном этапе обработки, мала. Это объясняется тем, что найденные значения параметров соответствуют малой окрестности минимума целевой функции, а задача минимизации, в свою очередь, является чаще всего корректно поставленной для выбираемых участков спектров.

В ходе разработки программы также были рассмотрены участки масс-спектров, содержащие пары пиков, плохо разрешённые настолько, что визуально трудно или невозможно выделить зоны соответствующие каждому пику. При обработке таких участков было определено, что используемая модель формы пика не подходит для анализа подобных участков, что связано с действительной несимметричностью формы пика. В связи с этим существенным препятствием необходимо ввести учёт реальной формы пиков. Для этого можно использовать более совершенные модели формы пиков или ввести учёт аппаратной функции.

По всей видимости, задача определения количества пиков на произвольно выбранном участке спектра неразрешима теми методами, которые возможно осуществить в рамках решения задачи анализа масс-спектров. Это замечание особенно существенно при рассмотрении спектров, содержащих не полностью разрешённые пики. Поэтому степень автоматизации, достигнутая в программе обработки масс-спектров, описанной в [6], при которой без вмешательства оператора каждому разрешённому пику ставится в соответствие собственная зона, не была достигнута в данной программе. Для успешного определения начальных значений параметров пика необходимо задание соответственной зоны, а для пары не полностью разрешённых пиков – зоны именно этой пары, ведь каждый пик выбранной пары можно считать парным для ещё одного – соседнего. Алгоритмически это можно осуществить методом перебора, но критерия правильности выбора всех произвольных пар не существует. Таким образом, процедура определения зон выполняется оператором.

При отладке программы было произведено сравнение результатов обработки тестовых масс-спектров разработанной программы bf с результатами программы TableCurve 2D. Результаты тестирования одного из спектров, рассмотренного в [6], приведённые в таблице совпали в пределах СКО, значения СКО совпали до третьей значащей цифры.

Совпадение результатов обработки выполняется и для множества других спектров, что свидетельствует о работоспособности и адекватности алгоритма программы bf.

На рис.1 изображён график одного из масс-спектров, использованных для тестирования программы. На масс-спектре видна пара не полностью разрешённых пиков, расположенных вблизи 27 а.е.м. Не полностью разрешённые пики интерпретированы как CHN (27,0109 а.е.м.) и C_2H_3 (27,02349 а.е.м.). На рис.2 изображено разложение на составляющие суммарного спектра этих пиков.

Преимуществом данной программы перед такими неспециализированными для решаемой задачи программными продуктами, как MathCAD, MatLab и TableCurve 2D, является высокая степень автоматизации вычислений на всех этапах обработки масс-спектра,

реализованных на данный момент. Программа имеет удобный интерфейс пользователя. На обработку участка масс-спектра, приведённого на рис. 1, затрачено время порядка 20 секунд. Это связано с тем, что подбор начальных параметров поиска выполняется автоматически.

Таблица

Результаты тестирования программ обработки масс-спектров.

Величина	TableCurve 2D		bf	
	Значение	СКО	Значение	СКО
n , ион/с	9,1	1,7(796)	9,1	1,7(796)
a_1, ион/с	20087	152	20088	152
r_1	1392,1	9,5	1392,3	9,5
m_1, а.е.м.	27,012607	0,000079	27,012606	0,000079
a_2, ион/с	3953	175	3955	175
r_2	3817	174	3813	173
m_2, а.е.м.	27,02922	0,000162	27,02922	0,000162

Примечание: в скобках указаны дополнительные совпавшие знаки.

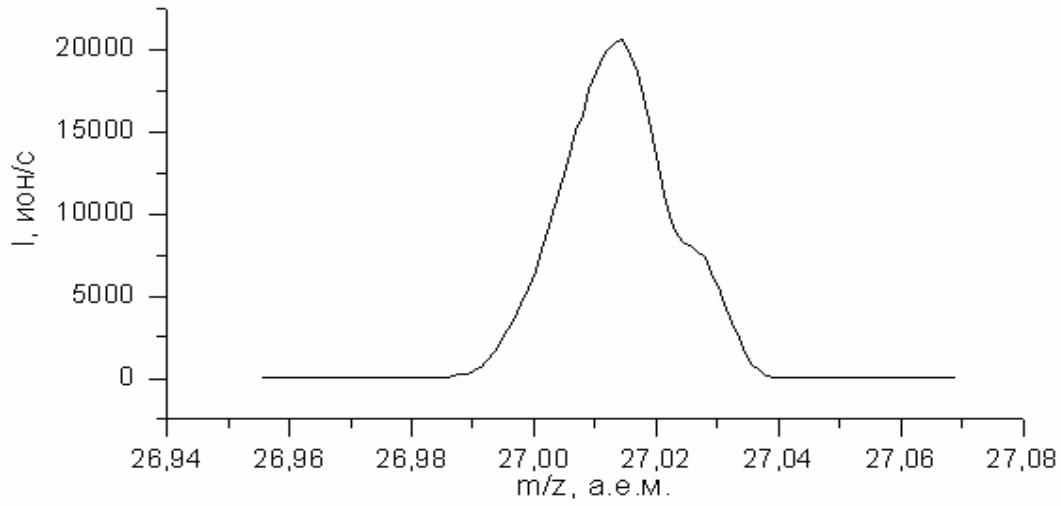

Рис.1. Не полностью разрешённые два пика, расположенные вблизи 27 а.е.м.

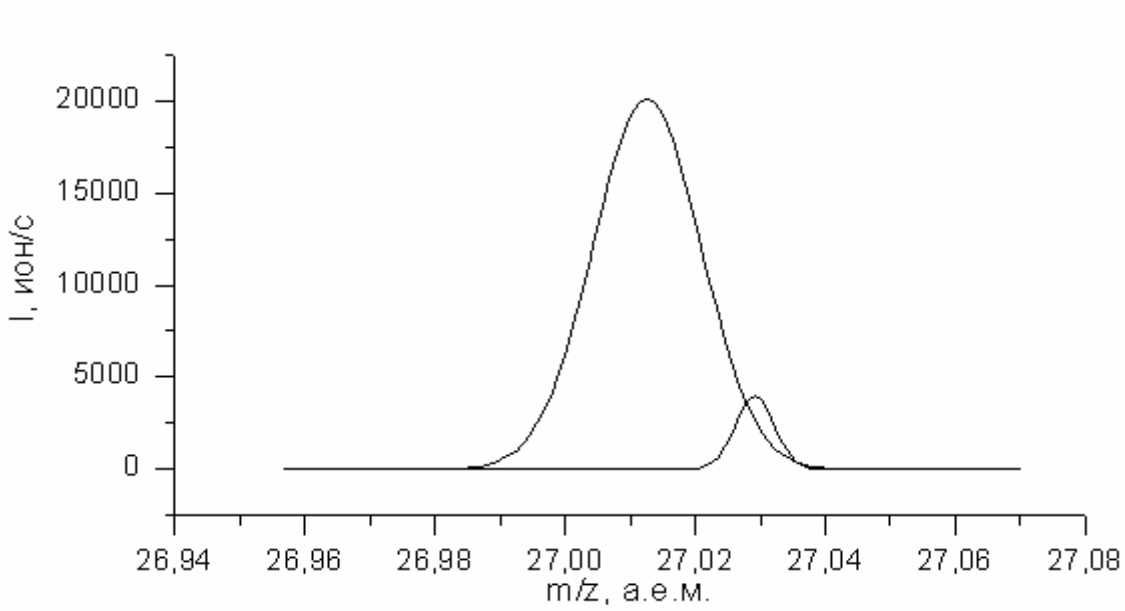

Рис. 2. Разложение на составляющие пика вблизи 27 а.е.м.

В основе программы лежит научно обоснованный метод получения оценок интересующих величин, учтён особый характер ионной статистики, результатом чего явилось повышение чувствительности общего метода. Численные эксперименты в ходе разработки и отладки программы показали, что в ряде случаев необходимо учитывать действительную несимметричность формы пиков на масс-спектре. Несимметричность формы пиков является одним из ограничивающих возможности общего метода программы анализа масс-спектров факторов. Планируется, что в ходе дальнейших исследований будут найдены способы учёта действительной формы пиков и программа будет дополнена средствами для их реализации.

ЛИТЕРАТУРА

1. Свиридас С.И. Обнаружение, разделение и оценка параметров масс-спектрометрических пиков методом свёртки экспериментальных данных с производными гауссовых функций / С.И. Свиридас, И.В. Заруцкий, А.М. Ларионанов, В.В. Манойлов // Научное приборостроение, 1999. Т. 9, № 2. С. 71-75.
2. Заруцкий И.В. Алгоритмы и программы первичной обработки масс-спектрометрических сигналов для автоматизации изотопного и элементного анализа / Автореф. дисс. … канд. техн. наук, Санкт-Петербург, 2007. 16 с.
3. Разников В.В. Анализ неполностью разрешенных масс-спектрометрических данных / В.В. Разников, А.Р. Пихтелев, М.О. Разникова // Масс-спектрометрия, 2006. Т. 3, № 2. С. 113-130.
4. Новиков Л.В. Автоматизированный измерительно-вычислительный комплекс специализированного масс-спектрометра МТИ-350Г / Автореф. дисс. … канд. техн. наук, Санкт-Петербург, 2006. 17 с.
5. Щиголев Б.М. Математическая обработка наблюдений. М.: Наука, 1969. 344 с.
6. Соломеин А.А. Методика калибровки массовой шкалы и расшифровки масс-спектра остаточных газов масс-спектрометра МИ-1201ФГМ / А.А. Соломеин, Б.А. Калинин, П.М. Глинских // Аналитика и контроль. 2006. Т. 10, № 1. С. 45-48.

MATHEMATICAL PROCESSING OF MASS-SPECTRUM THAT CONTAINS INCOMPLETE SEPARATED PEAKS

A.V. Tomilov, B.A. Kalinin, O.E. Aleksandrov, V.D. Seleznyov

The program has been developed for mathematical treatment of mass-spectrum. The program mean for treatment of single peak and incomplete separated peaks pair. The original algorithm of the treatment is proposed.

Key words: sector field mass analyzer, incomplete separated peaks, mass resolving, software assisted interpretation of mass spectra, peak shape model, peak shape approximation