

ОПТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 681.32 + 543.5 : 543.8

ОБНАРУЖЕНИЕ МАСС-СПЕКТРАЛЬНЫХ ПИКОВ
В БИОПРОБАХ ПРИ ДОПИНГОВОМ КОНТРОЛЕ

А. Г. Вострецов¹, В. А. Богданович², В. И. Будь³

¹Новосибирский государственный технический университет, г. Новосибирск
E-mail: vostretsov@adm.nstu.ru

²Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»,
г. Санкт-Петербург

³Национальный антидопинговый центр, г. Киев, Украина

Рассмотрено решение задач обработки масс-спектрометрических данных применительно к задачам допингового контроля. Показано, что такие задачи по результатам масс-спектрометрического анализа биопроб характеризуются высоким уровнем априорной неопределенности. Предложено процедуру обнаружения масс-спектральных пиков осуществлять на основе метода контраста, суть которого состоит в том, что сначала берется выборка из процесса, где полезный сигнал заведомо отсутствует (т. е. в отсутствие анализируемой пробы в рабочей камере масс-спектрометра, в этом случае регистрируемый масс-спектр будет зависеть лишь от характеристик измерительного канала и остатков примесей в рабочей камере), а затем – рабочая выборка, в которой необходимо обнаружить полезный сигнал. На основе статистических принципов инвариантности и несмещенности получены оптимальные алгоритмы обнаружения, приведены результаты статистических испытаний алгоритма и натурального эксперимента.

Введение. Допинговый контроль стал неотъемлемой частью национальных и международных соревнований, а также подготовки спортсменов к этим соревнованиям. Наиболее достоверные результаты контроля достигнуты при использовании масс-спектрометрических методов. Стремление спортсменов скрыть факты наличия допинга приводит к разработке различных схем применения запрещенных препаратов, чтобы их концентрация в анализируемых пробах была ниже уровня чувствительности аппаратуры для допингового контроля. Поэтому проблема повышения чувствительности используемых масс-спектрометров является первой и наиболее актуальной. Ее решение носит комплексный характер: это и совершенствование аппаратной части, и разработка новых методик, и применение новых методов регистра-

ции и обработки масс-спектрометрических данных. Особое место среди них занимают методы регистрации и обработки масс-спектрометрических данных. Правильный выбор этих методов позволяет обеспечить минимальные потери полезной информации и повысить достоверность получаемого результата.

Вторая проблема связана с большим объемом и сложностью интерпретации полученной информации в ходе масс-спектрометрических исследований биопроб, наличием шумовой составляющей, априорной неопределенностью статистических свойств зарегистрированных данных. Именно поэтому важная роль при выполнении масс-спектрометрического анализа отводится оператору – конечный результат во многом зависит от его квалификации. Перечисленные выше проблемы обуславливают высокие требования, предъявляемые к методам обработки информации, которые должны минимизировать потери полезной информации, обеспечить устойчивость характеристик контроля в условиях изменяющихся свойств регистрируемых данных и параметров аппаратуры, гарантировать высокую точность измерений и достоверность принимаемых решений. Кроме того, применяемые методы обработки информации должны способствовать автоматизации эксперимента вплоть до получения результата в автоматическом режиме. Последнее позволит исключить субъективный фактор в принятии решений и повысить их достоверность, что особенно важно при проведении допингового контроля.

Осознание необходимости использования сложных методов обработки в масс-спектрометрии привело к возникновению нового направления – информационно-аналитической масс-спектрометрии. Термин этот впервые предложен в монографии [1], которая была, по существу, первой, где системно рассмотрены основные аспекты нового направления.

Методы обработки масс-спектрометрических данных [1] основаны на использовании методов математической статистики и теории распознавания образов и ориентированы на довольно большой объем априорной информации. Вопросы устойчивости характеристик контроля в условиях изменяющихся свойств регистрируемых данных практически не рассмотрены. В представленной работе предлагается решение задачи обнаружения масс-спектральных пиков в условиях изменяющихся свойств регистрируемых данных применительно к допинговому контролю.

Первичная обработка масс-спектрометрических данных. В работе [1] отмечается, что во многих масс-спектрометрах для сжатия входной информации используется процедура первичной обработки – пороговое обнаружение масс-спектральных пиков. Система регистрирует только отсчеты, превысившие некоторый порог, при этом объем информации по опытным данным сокращается примерно в 50 раз. Недостатки этой процедуры: обнаружение ложных пиков, пропуск истинных пиков, их дробление и потери точности определения основных параметров пиков. Все перечисленные недостатки будут присущи любому алгоритму обнаружения в силу статистической природы наблюдаемого процесса. Поэтому речь может идти лишь о синтезе алгоритмов, которые обеспечили бы заданные статистические характеристики. Например, если обеспечить независимость вероятности ложного обнаружения масс-спектральных пиков от интенсивности фона, обусловленного шумами вторичного электронного умножителя, усилителя и сторонними примесями в анализируемой пробе, то можно гарантировать уровень достоверности получаемых результатов. Кроме того, желательным свойством алгоритмов обнаружения является минимизация вероятности

пропуска истинного пика (максимизация вероятности правильного обнаружения истинного пика). Перечисленным условиям отвечают равномерно наиболее мощные (РНМ) несмещенные алгоритмы, широко используемые в теории обнаружения сигналов [2]. Очевидно, что процедура обнаружения на основе сравнения отсчетов с фиксированным порогом в условиях изменяющихся характеристик фона не может обеспечить постоянство вероятности ложного обнаружения пика. Последняя будет зависеть, в частности, от дисперсии фона, что и имеет место на практике.

Рассмотрим, как можно модифицировать процедуру первичной обработки. Стабилизировать вероятность ложного принятия решения о регистрации отсчета при однократном измерении невозможно, следовательно, нужна дополнительная информация, в частности, о статистических свойствах фона. В теории обнаружения сигналов для получения информации о статистических свойствах шума широко используется метод контраста [2]. Суть его состоит в том, что сначала берется выборка из процесса, где полезный сигнал заведомо отсутствует, а затем – рабочая выборка, в которой необходимо обнаружить сигнал. Применительно к масс-спектрометрии это означает, что сначала нужно получить отсчеты в отсутствие анализируемой пробы в рабочей камере (в этом случае регистрируемый масс-спектр будет зависеть лишь от характеристик измерительного канала и остатков примесей в рабочей камере), а затем – при наличии анализируемой пробы. Первую выборку назовем опорной, вторую – рабочей. Рассмотрим, как будет в этом случае выглядеть алгоритм обнаружения.

Пусть в качестве АЦП используется счетчик ионов. Обозначим через x_i значение числа ионов, зарегистрированных в течение i -го интервала дискретизации длительностью τ масс-спектрограммы в отсутствие пробы анализируемого вещества, через y_i – при ее наличии. Отсчеты x_i и y_i распределены по закону Пуассона [1] и имеют совместное распределение вероятностей

$$p(x_i, y_i) = \frac{\exp\{-(\lambda_0 + \lambda_1)\}}{x_i! y_i!} \exp\{y_i \log(\lambda_1/\lambda_0) + (x_i + y_i) \log \lambda_0\}. \quad (1)$$

Здесь λ_0 и λ_1 – заранее неизвестные математические ожидания числа ионов, приходящих в течение i -го интервала дискретизации для опорной и рабочей выборок соответственно.

Распределение (1) характеризуется одномерным полезным параметром $\vartheta = \log(\lambda_1/\lambda_0)$ и одномерным мешающим параметром $\mu = \log(\lambda_0)$. Задача обнаружения сигнальной составляющей может быть сформулирована как задача проверки сложной статистической гипотезы относительно параметров распределения вероятностей (1):

$$H_0: \vartheta \leq 0, \quad \mu \in (-\infty, +\infty) \quad (\text{полезной составляющей в отсчете нет}); \quad (2)$$

$$H_1: \vartheta > 0, \quad \mu \in (-\infty, +\infty) \quad (\text{полезная составляющая в отсчете есть}).$$

Распределение (2) принадлежит экспоненциальному семейству, обладает достаточными статистиками $U = y_i$ и $T = x_i + y_i$, при гипотезе H_0 зависит лишь от параметра μ , область значений которого содержит одномерный ин-

тервал, поэтому по теореме о полноте [3] при гипотезе H_0 оно является полным и РНМ несмещенный алгоритм обнаружения примет вид [2]

$$\varphi(U, T) = \begin{cases} 1, & U > C(\alpha, T), \\ \psi, & U = C(\alpha, T), \\ 0, & U < C(\alpha, T). \end{cases} \quad (3)$$

Здесь $C(\alpha, T)$ – пороговая функция, зависящая от уровня вероятности ложного обнаружения α и достаточной статистики T ; $0 \leq \psi \leq 1$ – параметр рандомизации. Пороговая функция $C(\alpha, T)$ и константа ψ определяются из уравнения

$$\mathbf{M}[\varphi(U, T) | \mathfrak{G} = 0, T] = \alpha, \quad (4)$$

где $\mathbf{M}[\cdot | \mathfrak{G} = 0, T]$ – условное математическое ожидание при $\mathfrak{G} = 0$.

В работе [2] показано, что РНМ несмещенный алгоритм обнаружения (3), выраженный через исходные наблюдения, имеет вид

$$\varphi(x_i, y_i) = \begin{cases} 1, & y_i > C(\alpha, x_i + y_i), \\ \psi, & y_i = C(\alpha, x_i + y_i), \\ 0, & y_i < C(\alpha, x_i + y_i). \end{cases} \quad (5)$$

Пороговая функция $C(\alpha, x_i + y_i)$ и параметр рандомизации ψ зависят от заданного уровня вероятности ложного обнаружения α , значения суммы $x_i + y_i$ и определяются как решение уравнения (4), которое в данном случае принимает следующий вид:

$$\sum_{k=C(\alpha, x_i + y_i) + 1}^{x_i + y_i} \left[\binom{x_i + y_i}{k} \left(\frac{1}{2}\right)^{x_i + y_i} \right] + \psi \binom{x_i + y_i}{C(\alpha, x_i + y_i)} \left(\frac{1}{2}\right)^{x_i + y_i} = \alpha. \quad (6)$$

Если суммарное число ионов $x_i + y_i > 36$, то для вычисления пороговой константы $C(\alpha, x_i + y_i)$ и параметра рандомизации ψ можно воспользоваться аппроксимацией биномиального распределения (6) гауссовым распределением вероятностей со средним $\frac{x_i + y_i}{2}$ и дисперсией $\frac{x_i + y_i}{4}$. Соответствующие формулы приведены в [2].

На рис. 1 показаны зависимости вероятности правильного обнаружения β алгоритма (5) от отношения сигнал/шум $q = \lambda_1 / \lambda_0$ при различных значениях вероятности ложного обнаружения $\alpha = 0,01$ (рис. 1, *a*) и $\alpha = 0,001$ (рис. 1, *b*) и интенсивности шумовой компоненты λ_0 . Зависимости получены методом компьютерного моделирования алгоритма (5).

Из рисунка следует, что вероятность правильного обнаружения алгоритма (5) зависит не только от отношения сигнал/шум, но и от абсолютного значения математического ожидания λ_0 числа ионов в отсчетах опорной (шумовой) выборки. Если в рабочей выборке среднее значение числа ионов $\lambda_1 = 60$, то надежное обнаружение будет происходить только в том случае, когда сред-

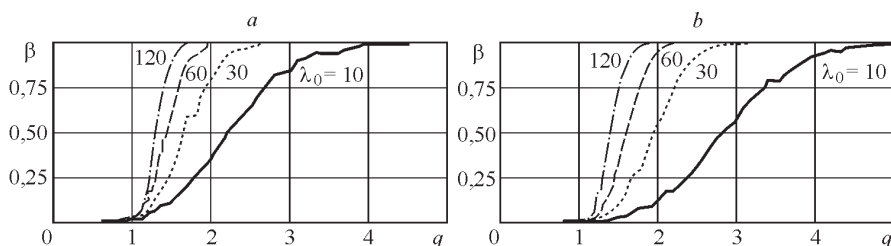


Рис. 1

нее значение числа ионов в шумовой выборке будет не более 30. Вероятность ложного обнаружения (на рис. 1 область $q \leq 1$) при этом не превысит заданный уровень α .

Предложенный алгоритм обнаружения обладает практически важными свойствами:

1. Вероятность ложного обнаружения не превышает заданное значение α при любых фактических значениях математического ожидания числа ионов в отсчетах рабочей выборки. Следовательно, алгоритм автоматически подстраивается под характеристики регистрируемого процесса.

2. Структура решающей функции алгоритма не зависит от фактических значений λ_0, λ_1 , т. е. они структурно устойчивы. Это свойство особенно важно при реализации в автоматических системах.

3. Вероятность правильного обнаружения алгоритма максимальна в классе всех несмещенных алгоритмов при любых значениях параметра q .

Обнаружение масс-спектрального пика в целом (совмещение первичной и вторичной обработки масс-спектрометрических данных). Особенность задачи допингового контроля состоит в том, что заранее известно, какие вещества следует обнаруживать в исследуемой пробе, неизвестными являются лишь сам факт их присутствия в пробе и концентрация. Поэтому временное положение и форма масс-спектральных пиков, соответствующих этим веществам, заранее известны, а неизвестными являются факт наличия соответствующих пиков в наблюдаемых данных и их площади. Учитывая эту особенность, задачу обнаружения масс-спектральных пиков можно сформулировать как задачу обнаружения пика в целом. Рассмотрим ее на примере обнаружения одного пика.

Как и прежде, для обеспечения устойчивости работы алгоритма обнаружения в условиях априорной неопределенности воспользуемся методом контраста. Обозначим через $\mathbf{x} = \{x_i, i=1, \dots, n\}$ выборочный вектор, составленный из значений числа ионов, зарегистрированных в течение n интервалов дискретизации длительностью τ масс-спектрограммы в отсутствие пробы анализируемого вещества, а через $\mathbf{y} = \{y_i, i=1, \dots, n\}$ – при ее наличии. Масс-спектрограмму фона будем задавать функцией $N\lambda(i)$, где $\lambda(i)$ – нормированная функция, определяющая ее форму, N – площадь под кривой спектрограммы (в общем случае N и $\lambda(i)$ заранее неизвестны). Число интервалов дискретизации n выберем таким образом, чтобы длительность анализируемого интервала спектрограммы $T = n\tau$ совпадала с длительностью обнаруживаемого пика. Форма пика задается функцией $Sf(i)$, где S – площадь пика, $f(i)$ – нормированная функция, определяющая форму пика (в большинстве случаев форма масс-спектральных пиков хорошо описывается гауссовой кривой

[1]). При наличии обнаруживаемого вещества в пробе $S > 0$, в отсутствие – $S = 0$. Совместное распределение вероятностей с учетом того, что отсчеты независимы и распределены по закону Пуассона, примет следующий вид:

$$p(\mathbf{x}, \mathbf{y}) = \exp \left\{ - \left[\sum_{i=1}^n (2N\lambda(i) + Sf(i)) \right] \right\} / \prod_{i=1}^n (x_i! y_i!) \times \\ \times \exp \left\{ \sum_{i=1}^n \log(N\lambda(i)) x_i + \sum_{i=1}^n \log(N\lambda(i) + Sf(i)) y_i \right\}. \quad (7)$$

Структура распределения вероятностей (7) не позволяет напрямую воспользоваться методами теории устойчивого обнаружения, так как полезный и мешающий параметры не разделены. Однако можно найти приближенное решение. Полагая $x_i \gg 1$, $y_i \gg 1$ и обозначив $X_i = 2\sqrt{x_i + 1}$, $Y_i = 2\sqrt{y_i + 1}$, $i = 1, \dots, n$, воспользуемся гауссовой аппроксимацией распределения $w(\mathbf{X}, \mathbf{Y})$ векторов $\mathbf{X} = \{X_1, \dots, X_n\}$, $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ и получим [4]

$$w(\mathbf{X}, \mathbf{Y}) = (1/2\pi)^n \times \\ \times \exp \left\{ - \frac{1}{2} \left[\sum_{i=1}^n (X_i - 2\sqrt{N\lambda(i)})^2 + \sum_{i=1}^n \left(Y_i - 2\sqrt{N\lambda(i) \left(1 + \frac{Sf(i)}{N\lambda(i)} \right)} \right)^2 \right] \right\}. \quad (8)$$

При $\max_i \left(\frac{Sf(i)}{N\lambda(i)} \right) \ll 1$, что соответствует наиболее интересному для практики случаю обнаружения слабого пика, выражение (8) упрощается:

$$w(\mathbf{X}, \mathbf{Y}) \approx (1/2\pi)^n \times \\ \times \exp \left\{ - \frac{1}{2} \left[\sum_{i=1}^n (X_i - 2\sqrt{N\lambda(i)})^2 + \sum_{i=1}^n \left(Y_i - 2\sqrt{N\lambda(i)} - \frac{Sf(i)}{\sqrt{N\lambda(i)}} \right)^2 \right] \right\}. \quad (9)$$

Распределение (9) симметрично относительно группы \mathbf{G} аддитивных преобразований. В отсутствие полезного сигнала ($S = 0$) группе \mathbf{G} соответствует произвольное изменение уровня фона. Максимальным инвариантом группы \mathbf{G} будет статистика $\mathbf{Z} = \{Z_1, \dots, Z_n\}$, $Z_i = (Y_i - X_i)/\sqrt{2}$, $i = 1, \dots, n$, [2] с распределением вероятностей

$$w(\mathbf{Z}) = (1/2\pi)^n \exp \left\{ - \frac{1}{2} \sum_{i=1}^n \left(Z_i - \frac{Sf(i)}{\sqrt{2N\lambda(i)}} \right)^2 \right\}. \quad (10)$$

Распределение (10) характеризуется векторным полезным параметром $\Delta = \{\delta_1, \dots, \delta_n\}$, $\delta_i = \frac{S}{\sqrt{2N\lambda(i)}} = q \sqrt{\frac{N}{2\lambda(i)}}$ ($q = S/N$ – отношение площадей по-

лезного и фонового пиков, которое будем называть отношением сигнал/шум), не зависит от априори неизвестных параметров сигнала и шума и определяется только соотношениями δ_i между уровнями сигнала и фона наблюдаемых отсчетов. При $S = 0$ (обнаруживаемое вещество в пробе отсутствует) распределение (10) не зависит ни от параметров фона, ни от параметров полезного сигнала. Это позволяет стабилизировать вероятность ложного обнаружения пика. Анализ выражения (10) показывает, что в общем случае ($n > 1$) не существует РНМ инвариантного алгоритма обнаружения, так как семейство распределений не обладает монотонным отношением правдоподобия. Однако можно построить алгоритм обнаружения с максиминными свойствами на естественном классе альтернатив [5], когда альтернатива H_1 состоит в том, что хотя бы одно из значений $\delta_i \geq \delta_i^* > 0$, $i = 1, \dots, n$, где δ_i^* – некоторое заданное число, определяющее границу «зоны безразличия». Такой алгоритм максимизирует минимальную вероятность правильного обнаружения. Так как распределение (10) удовлетворяет условиям теоремы 2 в [5], то решающая функция максиминного алгоритма примет вид

$$\varphi(\mathbf{Z}) = \begin{cases} 1, & \sum_{i=1}^n \exp \left\{ \delta_i^* f(i) Z_i - \frac{(\delta_i^* f(i))^2}{2} \right\} > C(\alpha), \\ 0, & \sum_{i=1}^n \exp \left\{ \delta_i^* f(i) Z_i - \frac{(\delta_i^* f(i))^2}{2} \right\} \leq C(\alpha). \end{cases} \quad (11)$$

Пороговая константа $C(\alpha)$ определяется заданным уровнем α ложного обнаружения пика.

Выражая (11) через исходные наблюдения, получим окончательный алгоритм обнаружения

$$\varphi(\mathbf{x}, \mathbf{y}) = \begin{cases} 1, & \sum_{i=1}^n \exp \left\{ \delta_i^* f(i) (\sqrt{2} \sqrt{y_i + 1} - \sqrt{2} \sqrt{x_i + 1}) - \frac{(\delta_i^* f(i))^2}{2} \right\} > C(\alpha), \\ 0, & \sum_{i=1}^n \exp \left\{ \delta_i^* f(i) (\sqrt{2} \sqrt{y_i + 1} - \sqrt{2} \sqrt{x_i + 1}) - \frac{(\delta_i^* f(i))^2}{2} \right\} \leq C(\alpha). \end{cases} \quad (12)$$

Анализ алгоритма (12) проводился методом имитационного моделирования на ЭВМ. Длительность обнаруживаемого пика была принята равной 21 отсчету, центр пика располагался на десятом отсчете, площадь фонового пика N изменялась от 10^2 до 10^4 , форма спектрограммы фона была принята равной суперпозиции постоянной составляющей и двух разнесенных между собой на 15 отсчетов отрезков гауссовых кривых (соотношение их максимальных значений составляло 0,1, 1,0 и 0,5). Отношение сигнал/шум q изменялось от 0 (когда сигнал отсутствовал) до 0,25 с шагом 0,005. Параметры δ_i^* , $i = 1, \dots, n$, были приняты одинаковыми во всех экспериментах и равными 2,5. Порог $C(\alpha) = 23,5$, при этом вероятность ложного обнаружения пика сохранялась на уровне порядка 10^{-3} во всех экспериментах.

На рис. 2, a пунктирной линией показана модель спектрограммы фона $N\lambda(i)$ ($N = 625$), сплошной – модель полезного пика $Sf(i)$ при отношении сигнал/шум $q = 0,25$. Соответствующие случайные реализации (пунктирная

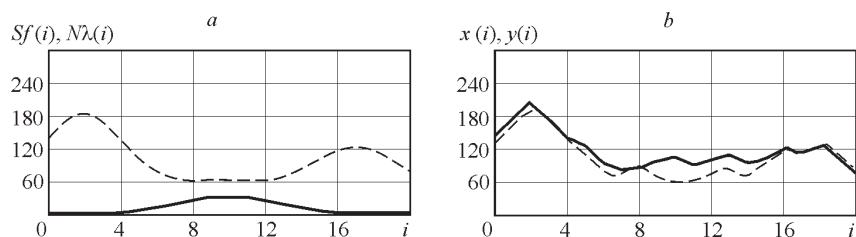


Рис. 2

линия – фоновая реализация, сплошная – реализация, содержащая полезный сигнал) приведены на рис. 2, *b*. Видно, что «на глаз» оператору принять решение о наличии пика затруднительно.

Зависимость вероятности правильного обнаружения пика от отношения сигнал/шум при различных значениях интенсивности фона показана на рис. 3 ($N = 300$ (*a*) и $N = 625$ (*b*)). Видно, что с ростом уровня фона вероятность правильного обнаружения растет. Это объясняется более богатой статистикой регистрируемых импульсов.

Еще раз отметим, что главной особенностью предложенного алгоритма является постоянство вероятности ложного обнаружения пика при любых уровнях и формах фоновой составляющей. Вероятность правильного обнаружения максимизируется в области малых значений отношений сигнал/шум и, как уже отмечалось, зависит от интенсивности и формы фоновой составляющей. Но даже для случая, когда в опорной выборке присутствует пик, совпадающий с полезным (наличие в камере масс-спектрометра загрязняющих примесей из ионов исследуемого вещества), вероятность ложного обнаружения пика остается приблизительно на прежнем уровне, что особенно важно при допинговом контроле, а отношение сигнал/шум, при котором будет обнаружено исследуемое вещество, увеличится незначительно. В ходе моделирования было установлено, что при наличии ионов исследуемого вещества в фоновой составляющей наблюдаемого процесса вероятность ложного обнаружения пика остается на том же уровне (10^{-3}), вероятность же правильного обнаружения достигает величины 0,98 при $q = 0,28$, а не при $q = 0,20$, как показано на рис. 3, *b*.

В заключение приведем фрагменты масс-спектрограмм, полученных с помощью масс-спектрометра фирмы “Varian” (рис. 4). Пунктирной линией показана спектрограмма $x(i)$ в отсутствие исследуемого вещества, сплош-

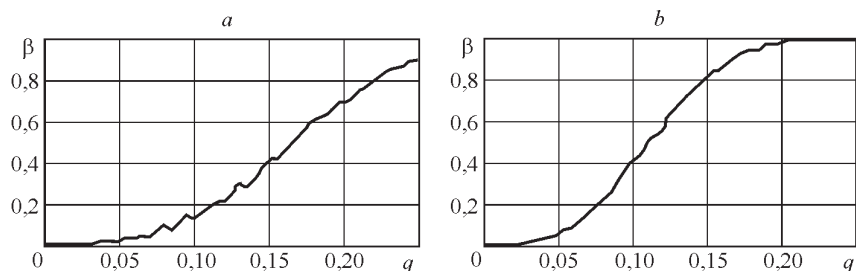


Рис. 3

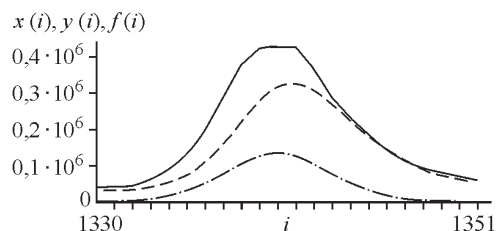


Рис. 4

ной – при его наличии ($y(i)$), штрихпунктирной – форма масс-спектрального пика $f(i)$, соответствующего обнаруживаемому веществу.

Спектрограммы были обработаны с помощью алгоритма (12), который обнаружил присутствие вещества. Хотя для доказательства эффективности предложенного алгоритма необходимо длительное экспериментальное исследование, результаты моделирования подтверждают перспективность его применения для допингового контроля.

Заключение. В представленной работе показано, что задачи информационно-аналитической масс-спектрометрии характеризуются высоким уровнем априорной неопределенности. На примере решения задач первичной обработки масс-спектрометрических данных и обнаружения заданного вещества по этим данным подтверждается плодотворность подходов, основанных на методах теории устойчивого обнаружения, различения и оценивания сигналов.

СПИСОК ЛИТЕРАТУРЫ

1. **Разников В. В., Разникова М. О.** Информационно-аналитическая масс-спектрометрия. М.: Наука, 1991.
2. **Богданович В. А., Вострецов А. Г.** Теория устойчивого обнаружения, различения и оценивания сигналов. М.: Физматлит, 2004.
3. **Леман Э.** Проверка статистических гипотез: Пер. с англ. Ю. В. Прохорова. М.: Наука, 1979.
4. **Левин Б. Р.** Теоретические основы статистической радиотехники. М.: Сов. радио, 1974. Кн. 1.
5. **Прокофьев В. Н.** Максимальное решение задачи обнаружения с векторным информативным параметром // Изв. АН СССР. Сер. Техническая кибернетика. 1978. № 5. С. 145.

Поступила в редакцию 31 января 2007 г.