

УДК 621.384.8:543.51

ДОСТОВЕРНОСТЬ МЕТОДОВ ИДЕНТИФИКАЦИИ В ХРОМАТО-МАСС-СПЕКТРОМЕТРИИ

А. Е. Малютин, к.ф.-м.н., доцент кафедры ОиЭФ РГРТУ, Рязань, Россия;
orcid.org/0000-0002-1704-0792, e-mail: aemalutin@mail.ru

***Целью работы** является исследование различных методов идентификации веществ в хромато-масс-спектрометрии и установление методов, оптимальных для автоматической идентификации токсичных веществ. Рассматриваются различные алгоритмы сравнения спектров в методе библиотечной идентификации, критерии принятия решения об идентификации, виды возникающих при этом ошибок и критерии достоверности идентификации. Производится сравнительное исследование различных алгоритмов идентификации малых количеств вещества с учетом факторов, искажающих масс-спектры, таких как статистический характер формирования масс-спектра, а также систематический и случайный фон. Определяются оптимальные критерии идентификации и пороги обнаружения веществ.*

***Ключевые слова:** хромато-масс-спектрометрия, масс-спектр, методы идентификации, достоверность идентификации, порог обнаружения вещества.*

DOI: 10.21667/1995-4565-2022-79-133-140

Введение

В настоящее время особую актуальность имеет задача разработки мобильных автоматических систем контроля веществ, представляющих угрозу для жизни человека [1, 2]. Такие системы в отличие от стационарных лабораторных хромато-масс-спектрометров особенно критичны к методам идентификации, так как необходимо достичь высокой достоверности без непосредственного участия высококвалифицированного исследователя. Основным методом идентификации в настоящее время является библиотечный поиск эталонного спектра, наиболее похожего на исследуемый спектр. В литературе встречается много различных алгоритмов, по которым проводится сравнение спектров [3-6], однако сравнение достоверности этих методов не производится – алгоритмы считаются равноценными. В то же время это сравнение является весьма актуальным, так как выбор алгоритма идентификации во многом определяет надежность распознавания.

Критерии сравнения спектров

Во всех методах для сопоставления спектров пробы и эталона рассчитывают количественные оценки степени совпадения (несовпадения) этих спектров в рамках выбранного алгоритма. К сожалению, даже терминология в этой области еще не устоялась: как синонимичные используются понятия «степень совпадения», «критерий подобия», «индекс подобия», «степень сходства», «мера близости», «фактор совпадения» и др. [3-6]. Часто величину этого параметра нормируют так, чтобы она находилась между 0 и 1. Существует большое количество алгоритмов вычисления этого параметра.

Достаточно часто в качестве меры близости выступает индекс взаимной корреляции. При этом спектры рассматриваются в виде векторов в многомерном пространстве. Массовые числа соответствуют декартовым осям координат в многомерном пространстве, а интенсивности – координатам по соответствующим осям. В таком случае угол между векторами изменяется от 0 до 90°, а косинус этого угла является индексом взаимной корреляции, изменяющимся от 1 для полного соответствия до 0 для полного несоответствия спектров. Обозначим $\vec{X} = (I_1^x, I_2^x, \dots, I_m^x)$ спектр определяемого неизвестного соединения (спектр пробы), а

$\vec{R} = (I_1^r, I_2^r, \dots, I_m^r)$ – библиотечный спектр некоторого соединения (эталонный или реперный спектр), тогда индекс взаимной корреляции будет находиться по формуле

$$C = \cos \alpha = \frac{\vec{X}\vec{R}}{|\vec{X}| |\vec{R}|} = \frac{\sum_i I_i^x I_i^r}{\sqrt{\sum_i (I_i^x)^2 \sum_i (I_i^r)^2}}. \quad (1)$$

Однако чаще в качестве параметра идентификации используется оценка несовпадения, которая равна 0 при полном совпадении и 1 при полном несовпадении, поэтому для сравнения будем использовать величину

$$D_1 = \sin^2 \alpha = 1 - \frac{(\vec{X}\vec{R})^2}{X^2 R^2} = 1 - \frac{\left(\sum_i I_i^x I_i^r\right)^2}{\sum_i (I_i^x)^2 \sum_i (I_i^r)^2}. \quad (2)$$

Вместо угла между векторами в многомерном пространстве можно использовать абсолютное (евклидово) расстояние, то есть модуль вектора разности

$$D_2 = |\vec{X} - \vec{R}| = \sqrt{\sum_i (I_i^x - I_i^r)^2}. \quad (3)$$

При этом спектры обычно нормируют так, что модуль вектора равен 1, то есть $\sum_i (I_i^x)^2 = 1$ и $\sum_i (I_i^r)^2 = 1$.

Иногда для ускорения расчетов используют более простую величину

$$D_3 = \sum_i |I_i^x - I_i^r|. \quad (4)$$

В этом случае нормировка проводится таким образом, чтобы сумма интенсивностей всех пиков в спектре была равна единице – $\sum_i I_i^x = 1$ и $\sum_i I_i^r = 1$.

В литературе встречается использование в такой формуле вместо интенсивности квадратного корня из интенсивности. При этом возрастает вклад малоинтенсивных пиков.

$$D_4 = \sum_i \left| \sqrt{I_i^x} - \sqrt{I_i^r} \right|, \quad (5)$$

где нормировка спектров – $\sum_i \sqrt{I_i^x} = 1$ и $\sum_i \sqrt{I_i^r} = 1$.

Учет большей значимости для идентификации более тяжелых массовых пиков производится умножением на массу иона

$$D_5 = \sum_i \left| m_i \sqrt{I_i^x} - m_i \sqrt{I_i^r} \right|, \quad (6)$$

где нормировка спектров – $\sum_i m_i \sqrt{I_i^x} = 1$ и $\sum_i m_i \sqrt{I_i^r} = 1$.

Наиболее сложный алгоритм расчета степени несовпадения спектров был предложен Мак-Лафферти:

$$D_6 = \sum_i \left(c I_i^r - I_i^x \right)^2, \quad (7)$$

где $c = \frac{\sum_i I_i^x I_i^r}{\sum_i (I_i^r)^2}$.

Критерии достоверности идентификации

При составлении ответа на запрос пользователя все отображенные в БД эталонные спектры ранжируются по величине параметров, которые в рамках принятого алгоритма характеризует степень их подобия спектру пробы. Пользователь получает перечень эталонов, занимающих первые места в ранжированном списке. При этом возможны два способа ограничения перечня: либо в машинный ответ включается постоянное и заранее заданное число эталонов, либо включаются все эталоны, у которых степень совпадения со спектром пробы оказалась выше определенного критического значения. Критерии отбора (критические значения ранжирующего параметра) обычно устанавливаются эмпирически.

Обычно окончательное заключение выносит сам исследователь, визуально сравнивая отображенные эталонные спектры со спектром пробы. Это связано с несовершенством компьютерных алгоритмов сопоставления спектров. К сожалению, они менее чувствительны к тонким различиям спектров, которые может принять во внимание исследователь. При работе в автоматическом режиме, когда за результат идентификации принимается эталон, стоящий во главе ранжированного списка, особое значение приобретает степень доверия полученным результатам.

Идентифицирующую способность чаще всего характеризуют величиной, равной отношению числа положительно решенных задач идентификации (N^+) к общему количеству контрольных поисков (N):

$$P = N^+ / N, \% . \quad (8)$$

Положительным решением считается то, в котором «неизвестное» соединение оказывается на первом месте машинного ответа. Эта величина зависит от выбора критерия отбора.

Возможны и другие приемы оценки результата идентификации. Например, в работах Мак-Лафферти для этой цели используется параметр

$$RL = 100 \frac{I_c}{I_c + I_f} .$$

Здесь RL (*reliability*) – достоверность принятия решения об идентификации объектов контрольной выборки; I_c и I_f – число корректных и ошибочных идентификаций при заданном значении параметра близости (различия) спектра запроса от спектров БД.

Результаты качественного анализа можно оценивать, пользуясь предложенными Мак-Лафферти идентификационными кривыми (*recall/reliability plots, RC/RL*). Алгоритм оценки не зависит от метода, объекта, поисковых признаков и критериев идентификации. Для этого используют большой массив результатов компьютерного качественного анализа объектов с известным составом. По оси абсцисс RC откладывают отношение числа верно опознанных веществ к общему числу опознававшихся компонентов, по ординате RL – отношение числа верно опознанных веществ к общему числу опознанных, в том числе и ложно идентифицированных. Применение менее строгих критериев опознания повышает результативность поиска, вероятность опознания растет, но при этом появляются ложные ответы, и однозначность результатов анализа снижается. Обработку массива статистических данных проводят при разных критериях идентификации, каждый режим дает точку на кривой $RL = f(RC)$.

Полученные кривые весьма наглядны, но, к сожалению, не дают возможности объективно и метрологически обоснованно оценивать, а тем более сравнивать надежность идентификации отдельных компонентов пробы. Как правило, потребителя анализа интересуют лишь некоторые компоненты пробы (например, опасные токсиканты), и ошибки в обнаружении этих компонентов не могут компенсироваться правильной идентификацией других компонентов даже в случае 99 % информативности анализа. Поэтому задача разработки априорных оценок надежности идентификации отдельных компонентов в хромато-масс-спектрометрическом анализе является чрезвычайно актуальной.

Для определения надежности идентификации можно рассмотреть два типа ошибок, которые могут произойти при принятии решения о присутствии компонента в пробе.

1. Отсутствующий компонент может быть признан присутствующим. Это – ложная идентификация (положительная ошибка, false positive, FP). Максимальная положительная ошибка будет наблюдаться при идентификации соединения, эталонный спектр которого в БД наибольшим образом коррелирует с эталонным спектром рассматриваемого соединения.

2. Присутствующий компонент пробы может быть признан отсутствующим. Это – пропуск сигнала (отрицательная ошибка, false negative, FN).

Для выбранного критерия отбора эти коэффициенты дают точку на плоскости FP/FN. Изменение величины критерия отбора приводит к построению на этой плоскости кривой, подобной кривым Мак-Лафферти, по которой можно визуальное оценить надежность идентификации. Как положительная, так и отрицательная ошибки могут изменяться в пределах от 0 до 1.

Основной задачей повышения степени достоверности идентификации является выбор такой величины критерия отбора, при которой обе эти ошибки наименьшие. Так как при ужесточении критерия отбора положительная ошибка уменьшается, а отрицательная растет, то в качестве целевой функции для оптимизации его значения можно использовать величину

$$K = (1 - FP)(1 - FN). \quad (9)$$

Сравнительное исследование алгоритмов идентификации

Идентифицирующие способности алгоритмов рассматривались по следующей методике. Масс-спектр выбранного целевого соединения предварительно модифицировался. При этом учитывались два фактора, приводящих к искажению спектра.

Во-первых, учитывался вероятностный характер процесса формирования масс-спектра. Фрагментные ионы образуются с вероятностью, которая определяет интенсивность соответствующих массовых пиков. Таким образом, статистические отклонения масс-спектра от эталонного можно получить, моделируя процесс формирования масс-спектра из заданного количества ионов. Для определения вероятности фрагментации используется эталонный спектр. При уменьшении числа ионов, создающих масс-спектр, отклонения от эталона будут значительно больше, что характерно для определения следовых количеств веществ.

Во-вторых, учитывалось влияние фона. Фон в масс-спектре может быть двух типов: систематический и случайный. Систематический фон вызван наличием спектра не разделенной хроматографически примеси или выносом жидкой фазы из колонки. Для удаления этого фона существуют довольно развитые математические методы. Случайный фон возникает, как правило, за счет шума системы регистрации. Моделировать подобный фон можно случайным распределением по равномерному закону, задавая число отсчетов (ионов) на 10 а.е.м. Анализ реальных спектров, полученных на монопольном масс-спектрометре, показал, что фон изменяется в пределах от 7 до 500 ион/дек. Эти величины и использовались при моделировании спектров.

Полученный модифицированный спектр целевого соединения сравнивался затем с эталонными спектрами БД по заданному алгоритму. Критерий отбора не задавался, а в качестве результата идентификации принимался эталон, для которого значение коэффициента было минимальным. В том случае, когда результатом оказывался эталон, из которого был получен модифицированный спектр, задача идентификации считалась решенной положительно. В качестве сравнительной характеристики использовалась идентифицирующая способность, рассчитанная по формуле (8). Идентифицирующая способность рассчитывалась для разного числа ионов анализируемого вещества и разного фона. По этим данным строились идентификационные кривые как зависимость идентифицирующей способности P от величины полезного сигнала $N_{\text{ион}}$ при заданном уровне фона.

Для проведения такого сравнительного анализа требуется выбрать библиотеку спектров и целевые соединения, для которых и будет проводиться анализ. В качестве библиотеки использовалась база данных масс-спектров Национального института стандартов и технологии США (NIST), находящаяся в свободном доступе. Эта БД содержит большое количество соединений и считается одной из самых надежных. В качестве целевых соединений были выбраны сильно-

действующие ядовитые и отравляющие вещества: аммиак, синильная кислота, несимметричный диметилгидразин, хлорциан, фосген, триэтиламин и 2,6-диметилбензоамин (ксилидин).

Результаты анализа приведены для фосгена, так как из рассматриваемых соединений фосген наименее специфичен. На рисунке 1 представлены кривые идентификационной способности без фона (а) и с фоном 10 ион/дек (б). Из приведенных графиков можно сделать следующие выводы. Во-первых, в отсутствие фона все параметры примерно равноценны, несколько более эффективно применение параметра D_5 . Во-вторых, с увеличением фона кривые идентификационной способности разбиваются на три группы: наихудшими характеристиками обладает параметр D_4 , наилучшими – D_1 , D_2 и D_6 , параметры D_3 и D_5 обладают средними характеристиками. Дальнейшее увеличение уровня фона не приводит к существенным расхождениям между коэффициентами D_1 , D_2 и D_6 . Это значит, что для дальнейшего анализа алгоритмов сравнения требуется рассмотреть какой-либо другой аспект.

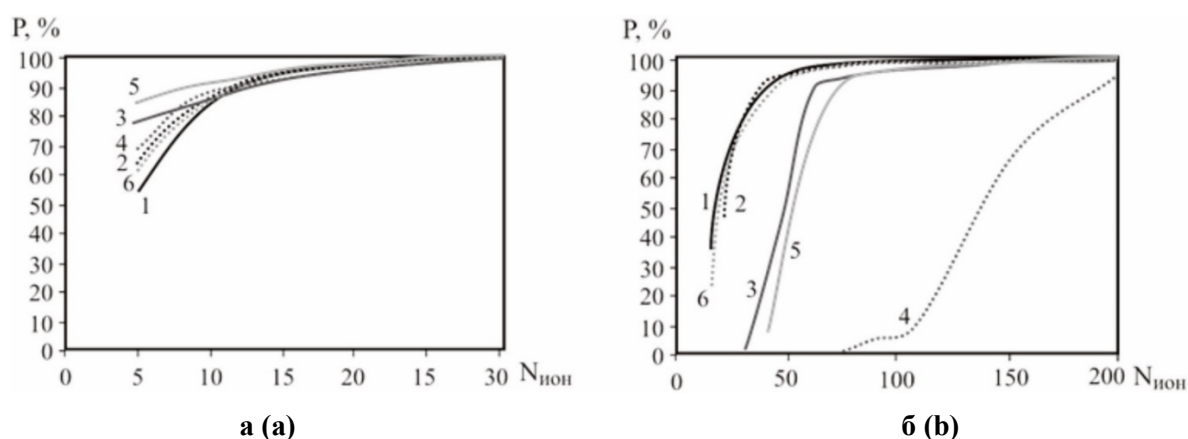


Рисунок 1 – Идентифицирующая способность для фосгена по критериям D_1 – D_6 в зависимости от числа ионов, образующих масс-спектр: а – при отсутствии фона; б – с фоном 10 ион/дек
Figure 1 – Identifying ability for phosgene according to criteria D_1 – D_6 , depending on the number of ions forming mass spectrum: a – in the absence of background and b – with background of 10 ions/dec

При увеличении фона для достижения заданного уровня идентифицирующей способности требуется увеличивать величину критерия отбора. Проведем для коэффициентов D_1 , D_2 и D_6 анализ относительного изменения критерия отбора D/D_0 , необходимого для достижения 95 %-го уровня идентифицирующей способности (D_0 – величина критерия отбора в отсутствие фона, D – величина критерия отбора при заданном фоне). Полученные результаты представлены на рисунке 2. Из них видно, что при изменении фона наиболее устойчивым оказывается коэффициент D_1 . Этот коэффициент и будем использовать для дальнейшего анализа достоверности идентификации.

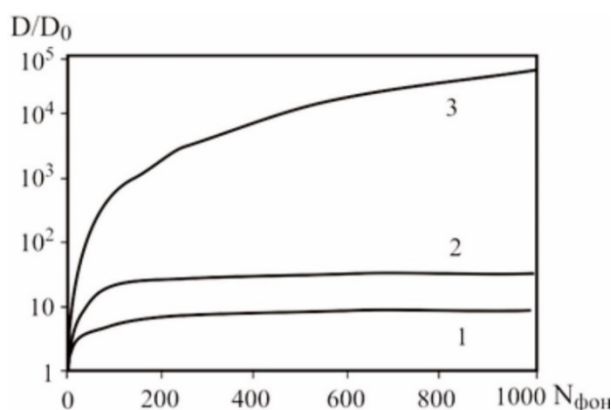


Рисунок 2 – Относительное изменение критерия отбора D/D_0 , соответствующего 95 %-му уровню идентифицирующей способности: 1 – D_1 , 2 – D_2 , 3 – D_6
Figure 2 – Relative change in selection criterion D/D_0 corresponding to 95 % level of identifying ability: 1 – D_1 , 2 – D_2 , 3 – D_6

Исследование достоверности идентификации

Для исследования достоверности идентификации воспользуемся изложенным выше методом. Расчет отрицательной ошибки будем проводить, идентифицируя при заданном критерии отбора модифицированный спектр фосгена. Отрицательной ошибкой FN будет являться отношение числа случаев, в которых фосген не будет обнаружен, к общему числу попыток. Положительной ошибкой FP будет являться отношение числа случаев, в которых модифицированный спектр этандиоида дихлорида при идентификации будет распознаваться как фосген.

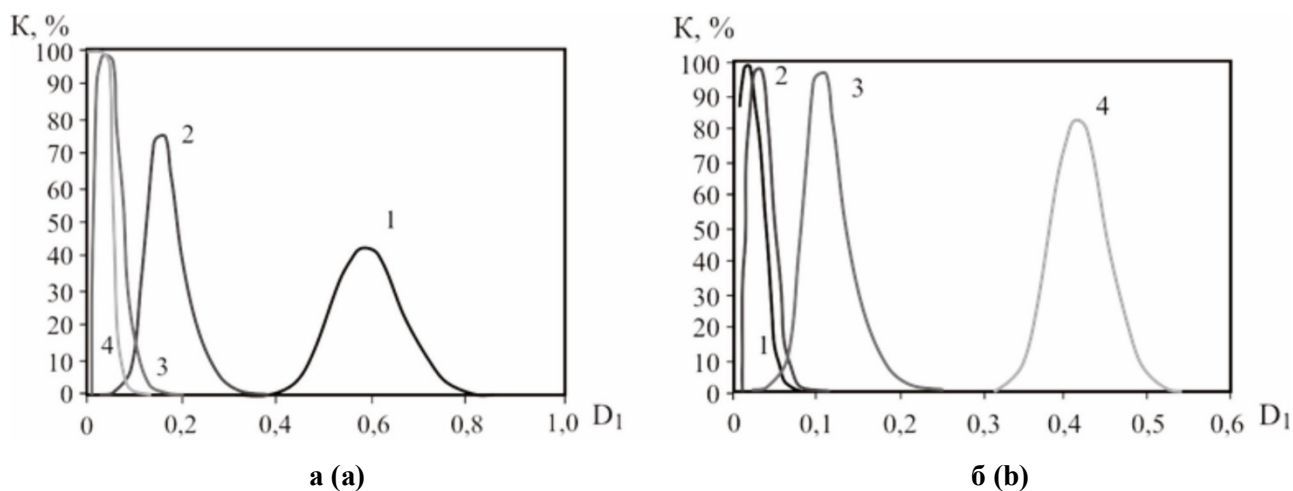


Рисунок 3 – Достоверность идентификации фосгена:

- а – для различного числа ионов, образующих спектр, при постоянном фоне;
 - б – для различных уровней фона при постоянном числе ионов, образующих спектр
- Figure 3 – Reliability of phosgene identification:
- a – for a different number of ions forming a spectrum with constant background;
 - b – for different background levels with a constant number of ions forming the spectrum

Выбор фосгена (карбонила дихлорида) и этандиоида дихлорида обусловлен малым различием их спектров. Различие спектров наблюдается только в области малоинтенсивных пиков, которые могут как отсутствовать вследствие вероятностного формирования спектра, так и быть скрыты фоном. При использовании коэффициента D_1 сравнение эталонных спектров дает величину, равную 0,056. Следовательно, использование этих веществ даст нижние оценки достоверности идентификации.

По результатам проведения численного эксперимента были построены идентификационные кривые на диаграмме FP/FN. По этим кривым можно для заданного уровня фона и числа ионов, формирующих спектр, оценить возможные значения положительной и отрицательной ошибок идентификации. Такой подход, предложенный Мак-Лафферти, дает наглядное представление о возможностях метода, но не позволяет получить оптимальное значение критерия отбора.

Для этого воспользуемся величиной K (9), представляющей собой степень достоверности идентификации соединения. На рисунке 3, а приведены зависимости этой величины от значения критерия отбора при постоянном значении фона (10 ион/дек) для различного числа ионов, образующих спектр (1 – 30 ионов, 2 – 100 ионов, 3 – 300 ионов, 4 – 1000 ионов). На рисунке 3, б приведены зависимости степени достоверности идентификации от значения критерия отбора для случая изменяющегося фона (1 – без фона, 2 – 10 ион/дек, 3 – 30 ион-дек, 4 – 100 ион/дек) при постоянном значении числа ионов, формирующих спектр (300 ионов).

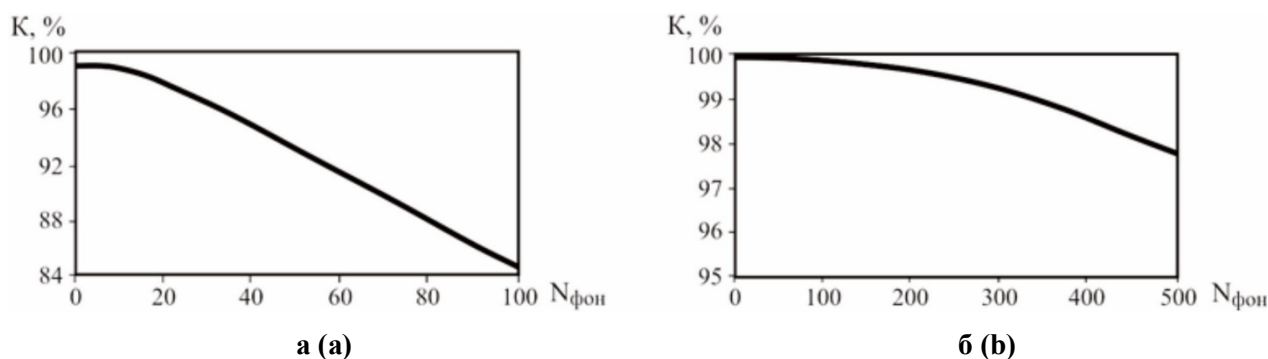


Рисунок 4 – Зависимость максимума достоверности идентификации фосгена от величины фона для спектров, образованных 300 ионами (а) и 1000 ионов (б)
Figure 4 – Dependence of maximum reliability of phosgene identification on background level for the spectra formed by 300 ions (a) and 1000 ions (b)

Заключение

На основании проведенных исследований можно сделать следующие выводы.

Во-первых, хотя в отсутствие фона идентифицирующая способность системы по фосгену может достигать 100 %, но при соответствующем значении критерия отбора возможна значительная вероятность положительной ошибки. Необходимое 95 %-е значение достоверности в отсутствие фона может быть достигнуто только для числа ионов, формирующих спектр, свыше 300. Эта величина представляет собой предел обнаружения фосгена.

Во-вторых, при увеличении полезного сигнала по сравнению с фоном происходит смещение максимума кривых достоверности в сторону меньшего значения критерия отбора. Причем ширина этих кривых получается довольно узкой, а это значит, что для автоматического распознавания веществ с высокой достоверностью величину критерия отбора нужно корректировать согласно величине сигнала и фона.

Зависимость максимумов степени достоверности идентификации от величины фона приведена на рисунке 4. По этим данным видны пределы обнаружения фосгена при различных уровнях фона. Так как максимальный уровень фона, оцененный по реальным спектрам, составлял величину порядка 500 ион/дек, то в самом худшем случае предел обнаружения фосгена при 95 % достоверности идентификации составляет 1000 ионов.

Библиографический список

1. **Экман Р.** Масс-спектрометрия: аппаратура, толкование и приложения / Экман Р., Зильберинг Е., Вестман-Бринкмальм Э., Край А. М.: Техносфера, 2013.
2. **Лебедев А. Т.** Масс-спектрометрия в органической химии. М.: Техносфера, 2015.
3. **Вершинин В. И.** Компьютерная идентификация органических соединений / Вершинин В. И., Дерендяев Б. Г., Лебедев К. С. М.: Наука, 2002.
4. **Мильман Б. Л.** Введение в химическую идентификацию. СПб.: ВВМ, 2008.
5. **Другов Ю. С.** Газохроматографическая идентификация загрязнений воздуха, воды, почвы и биосред / Ю. С. Другов, И. Г. Зенкевич, А. А. Родин М.: Лаборатория знаний, 2020.
6. **Мильман Б. Л., Журкович И. К.** Обобщенные критерии идентификации химических соединений методами хроматографии – масс-спектрометрии // Аналитика и контроль. 2020. Т. 24. № 3. С. 164-173.

UDC 621.384.8:543.51

RELIABILITY OF IDENTIFICATION METHODS IN CHROMATOGRAPHY-MASS SPECTROMETRY

A. E. Malyutin, Ph.D. (Phys. and Math.), associate professor, RSREU, Ryazan, Russia;
orcid.org/0000-0002-1704-0792, e-mail: aemalutin@mail.ru

The aim of the work is to study various methods for identifying substances in gas chromatography-mass spectrometry and to establish methods being optimal for automatic identification of toxic substances. Various algorithms for comparing spectra in library identification method, criteria for making a decision on identification, types of errors arising in this case, and criteria for identification reliability are considered. A comparative study of various algorithms for identifying small amounts of substances is carried out, taking into account the factors that distort mass spectra, such as statistical nature of mass spectrum formation as well as systematic and random background. Optimal identification criteria and detection thresholds for substances are determined.

Key words: chromatography-mass spectrometry, mass spectrum, identification methods, identification reliability, substance detection threshold.

DOI: 10.21667/1995-4565-2022-79-133-140

References

1. **Ekman R.** *Mass-spektrometriya: apparatura, tolkovanie i prilozheniya* (Mass Spectrometry: Instrumentation, Interpretation, and Applications). R. Ekman, E. Silbering, E. Westman-Brinkmalm, A. Kraj John Wiley & Son, 2009. (in Russian).
2. **Lebedev A. T.** *Mass-spektrometriya v organicheskoy himii* (Mass spectrometry in organic chemistry) Moscow.: TEKHNOSFERA, 2015. (in Russian).
3. **Vershinin V.** *Komp'yuternaya identifikaciya organicheskikh soedinenij* (Computer identification of organic compounds). Vershinin V. I., Derendyaev B. G., Lebedev K. S. Moscow: Nauka, 2002. (in Russian).
4. **Milman B. L.** *Vvedenie v himicheskuyu identifikaciju* (Introduction to chemical identification). St. Petersburg: VVM, 2008. (in Russian).
5. **Drugov Yu. S.** *Gazohromatograficheskaya identifikaciya zagryaznenij vozduha, vody, pochvy i biosred* (Gas chromatographic identification of air, water, soil and biological environment pollutants). Yu. S. Drugov, I. G. Zenkevich, A. A. Rodin. M.: Laboratory of Knowledge, 2020. (in Russian).
6. **Milman B. L., Zhurkovich I. K.** Obobshchennye kriterii identifikacii himicheskikh soedinenij metodami hromatografii – mass-spektrometrii. *Analitika i kontrol'*. 2020, vol. 24, no. 3, pp. 164-173 (in Russian).