Применение критерия Стьюдента для определения достоверности идентификации веществ при хроматографическом анализе

Ю.А. Кудеяров, Е.В. Кулябина, О.Л. Рутенберг

Всероссийский научно-исследовательский институт метрологической службы

Предложен метод оценки достоверности идентификации веществ при хроматографическом анализе, основанный на применении критерия Стьюдента в качестве статистического критерия для случая, когда для идентификации необходима дополнительная измерительная информация.

Ключевые слова: хроматографический анализ, идентификация, критерий Стьюдента.

Application of the Student criterion for identification at the chromatographic analysis

Yu. A. Kudeyarov, E.V. Kulyabina, O.L. Rutenberg

Method of substances identifying at chromatographic analysis based on the Student criterion is proposed for case when additional information is required.

Key words: chromatographic analysis, identification, Student criterion.

Распространенными методами определения состава веществ являются высокоэффективная жидкостная хроматография (ВЭЖХ), газо-жидкостная хроматография, масс-спектрометрия и т.п. При этом одной из основных задач, решаемых при хроматографическом анализе, является задача идентификации веществ в анализируемых пробах. Как по поставленной цели, так и по используемым методам достижения этой цели идентификация представляет собой частную задачу более общей проблемы распознавания образов, т.е. проблемы классификации и идентификации явлений, сигналов, объектов, которые характеризуются конечным набором свойств и признаков. Одним из самых распространенных способов, используемых при идентификации, в по-

следнее время стал способ, основанный на использовании для этих целей баз данных (БД). Таким образом, в случае анализа пробы, например, методом ВЭЖХ с применением БД вещество считается идентифицированным, если имеется однозначное совпадение характеристик анализируемого компонента с характеристиками одного из веществ БД.

Необходимо отметить, что одной из основных проблем идентификации является количественная оценка ее достоверности.

Идентификации органических соединений посвящено большое количество публикаций [1-4], во многих из них рассматриваются вопросы, связанные с достоверностью идентификации.

Так, в работе [1] указано, что однозначной идентификации соответствует ситуация, в которой одному аналитическому сигналу соответствует одно вещество, идентификация становится сложнее, когда сигналы одинаковы для нескольких веществ. В [1] подробно рассмотрен вероятностный подход определения достоверности идентификации, основанный на применении условных вероятностей и теоремы Байеса, определены вероятности ошибок первого и второго рода. Критерием идентификации являлся параметр удерживания. Недостатком работы является отсутствие примеров применения изложенного подхода к идентификации конкретных веществ, не ясна также связь предложенных теоретических положений с экспериментально измеренными характеристиками веществ.

В работе [2] рассмотрен подход к идентификации, основанный на применении так называемого значения длины списка. При таком подходе в систематическом токсикологическом анализе определяют характеристики удерживания (относительное время удерживания) желательно с использованием больше чем одной аналитической системы. Затем применяют «метод окна», в котором «окно» устанавливают для неизвестного компонента. Если вещество из БД попадает внутрь «окна», то оно становится возможным кандидатом на роль неизвестного вещества. Далее составляют список всех возможных кандидатов. Однако возможны ситуации, когда одни из кандидатов

находятся ближе к центру «окна», другие – у самого края, а третьи в одной аналитической системе попадают за пределы «окна», а в другой – внутрь. Для принятия решения об однозначности идентификации во всех приведенных ситуациях оценивают достоверность идентификации каждого кандидата. Далее составляют список всех кандидатов по убыванию достоверности и, применяя различные критерии отбора, выявляют наиболее предпочтительного кандидата. В принципе, описанная процедура является характерной для большинства способов идентификации.

В работах [3,4] рассмотрен подход, базирующийся на вычислении ошибок 1-го и 2-го рода при идентификации и эмпирическом выборе статистического критерия. В работе [4] содержится также анализ причин возможных ошибок идентификации. Отмечено, что ошибки при идентификации по одному параметру, а именно по времени удерживания, могут возникать по нескольким причинам. Наиболее характерными причинами являются:

1. случайный сдвиг пиков, который может происходить из-за колебаний скорости газа-носителя или скорости мобильной фазы при хроматографировании, а также из-за низкой повторяемости объема вводимой пробы и другим причинам;
2. невысокая межлабораторная воспроизводимость времени удерживания, поскольку значения времени удерживания в общем случае не являются постоянными величинами, а изменяются при переходе от анализа на одной колонке к анализу на другой;
3. случайные совпадения характеристик удерживания различных веществ и др.

В работе [4] при рассмотрении проблемы идентификации применительно к хроматографическому анализу авторы ограничиваются рассмотрением идентификации только по одной измеряемой характеристике – времени удерживания. При этом отмечено, что привлечение дополнительной информации о свойствах анализируемых проб повышает надежность и достовер-

ность идентификации. Методы реализации этого утверждения авторы не предлагают.

В настоящей статье описан метод, основанный на применении критерия Стьюдента, позволяющий провести идентификацию веществ при хроматографическом анализе и количественно оценить достоверность идентификации в случае, когда не удается провести идентификацию по объему удерживания, и возникает необходимость привлечения для этой цели дополнительной информации.

Прежде чем переходить к изложению решения поставленной задачи, необходимо определиться с основными понятиями, которые будут использоваться в дальнейшем.

При идентификации рассматривают данные двух типов.

Данные первого типа – данные из базы данных (БД-2012), полученные путем хроматографирования и записи спектра стандартных образцов состава чистых веществ или чистых веществ из надежных источников с содержанием основного вещества не менее 98 % и с известным содержанием примесей. БД-2012 содержит средние значения характеристик $a_R(i,j), \quad (i=1,2,...,m, \quad j=1,2,...,k)$, индекс i обозначает порядковый номер вещества в базе (их предполагается m), j - номер характеристики (их для каждого i-го вещества их предполагается k). В рассматриваемом случае характеристики веществ это объем удерживания и спектральные отношения, определяемые по экспериментально измеренным ультрафиолетовым спектрам (спектральным отношением является отношение значения оптической плотности на одной длине волны к значению оптической плотности того же вещества на другой длине волны). При этом каждой характеристике - объему удерживания и спектральным отношениям – соответствуют значения $\sigma_R(i,j)$ и допуски $d(i,j)$ на их значения. Рекомендуемый интервал, в котором должны находиться допуски [4], приведен ниже.

$$2\sigma_R(i,j) < d(i,j) < 3\sigma_R(i,j). \qquad (1)$$

Не все характеристики веществ в базе данных равноценны. Среди них необходимо выделить, в первую очередь, объем удерживания $V_R(i,1)$, по которому происходит основная идентификация, и которой в благоприятных случаях все может и ограничиться.

Данные второго типа это экспериментально измеренные характеристики веществ $x(i)$, такие как средние значения объемов удерживания, спектральные отношения $b_x(i,j)$ и соответствующие значения среднего квадратического отклонения $\sigma_x(i,j)$.

Для того чтобы иметь возможность применить статистические методы к идентификации, необходимо определить границы применимости теории.

Основное предположение сводится к тому, что значения характеристик веществ при измерениях ведут себя как случайные величины, а их распределение является нормальным. Однако следует признать, что сама характеристика анализируемого вещества (объем удерживания или спектральное отношение) не может быть случайной величиной. В то же время разность между временами удерживания или соответствующими спектральными отношениями на одной и той же длине волны у различных веществ уже будет случайной величиной, и к ней можно применять методы математической статистики.

Важным аспектом идентификации является выбор меры подобия или различия между измеренным параметром и значением, находящимся в базе данных.

Мерой различия для рассматриваемого случая является сдвиг (невязка) $\Delta_x(i,j)$ между измеренным средним значением характеристики (например, объемом удерживания) и значением этой характеристики в базе данных:

$$\Delta_x(i,j) = | a_R(i,j) - b_x(i,j) | . \qquad (2)$$

Идентификация считается результативной, если разность между измеренным значением характеристики и приведенным в БД будет не больше не-

которого критерия (в нашем случае – допуска на это значение), то есть если выполняется условие

$$\Delta_x(i, j) \le d(i, j). \qquad (3)$$

Условие (3) является необходимым, но недостаточным для однозначной идентификации.

При проведении идентификации формулируют статистические гипотезы H_0 (нулевая гипотеза) и H_1 (альтернативная гипотеза).

В нашем случае нулевая гипотеза H_0 заключается в утверждении, что вещество в пробе отсутствует.

Если при этом результат измерения показал наличие вещества в пробе, т.е. из-за погрешности измерений его результат удовлетворяет условию (3), и если на основании такого измерения принимается решение о наличии вещества в пробе, то совершается *ошибка 1-го рода* (ложноположительный результат). Другими словами, при ошибке 1-го рода гипотеза H_0 отвергается и принимается гипотеза H_1.

Вероятность ошибки 1-го рода обозначается α. Таким образом, α - это вероятность принятия альтернативной гипотезы, которая называется *уровнем значимости*.

Альтернативная ситуация заключается в справедливости гипотезы H_1.

Если при этом результат измерения показал отсутствие вещества в пробе, т.е. из-за погрешности измерений его результат не удовлетворяет условию (3), и если на основании такого измерения принимается решение об отсутствии вещества в пробе, то совершается *ошибка 2-го рода* (ложноотрицательный результат). Другими словами, при ошибке 2-го рода гипотеза H_1 отвергается и принимается гипотеза H_0. Вероятность ошибки 2го рода обозначается β.

Для наглядности, описанные выше ситуации можно представить в виде таблицы

Истинная ситуация	Результат идентификации	Характер ошибки
Определяемое вещество в пробе отсутствует (H_0)	Идентификация дает отрицательный результат	Ошибка отсутствует (гипотеза H_0 принимается)
Определяемое вещество в пробе отсутствует (H_0)	Идентификация дает положительный результат	Ошибка 1го рода (ложноположительный результат), гипотеза H_0 отвергается, принимается гипотеза H_1
Определяемое вещество в пробе присутствует (H_1)	Идентификация дает отрицательный результат	Ошибка 2го рода (ложноотрицательный результат), гипотеза H_1 отвергается, принимается гипотеза H_0
Определяемое вещество в пробе присутствует (H_1)	Идентификация дает положительный результат	Ошибка отсутствует (гипотеза H_1 принимается)

Достоверность идентификации можно описать функцией P, связанной с вероятностями α и β соотношением

$$P = 1 - \alpha - \beta. \qquad (4)$$

Функцию P нельзя полностью отождествлять с вероятностью правильной идентификации. На самом деле она может принимать значения от - 1 до + 1 в зависимости от значений α и β.

Отрицательные значения P получаются тогда, когда $\alpha + \beta > 1$, например, когда $\alpha = 0{,}5$, а $\beta = 0{,}6$.

Идеальная идентификация ($P = 1$) при этом отвечает случаю $\alpha + \beta = 0$, т.е. когда все ошибки идентификации отсутствуют, что невозможно.

Случай $P = 0$ (совершенно неверная идентификация) соответствует, например, ситуации, когда вещество в пробе отсутствует, а вероятность его ложной идентификации максимальна, либо когда $\alpha + \beta = 1$.

Только тогда, когда $0 < P < 1$, эту функцию условно можно отождествить с вероятностью правильной идентификации. Несмотря, на условный

характер этой функции, ее можно использовать в качестве количественной характеристики правильности идентификации.

Необходимо отметить, что приведенные выше понятия и определения известны достаточно хорошо и приведены, например, в работе [4].

Применим изложенную выше теорию к анализу конкретных экспериментальных данных.

Пусть в результате хроматографического эксперимента для трех веществ получены следующие значения объемов удерживания

$$V(1,1) = 3302 \text{ мкл}, \quad V(2,1) = 3569 \text{ мкл}, \quad V(3,1) = 3304 \text{мкл}. \qquad (5)$$

Сравнивая эти значения с соответствующими значениями объемов удерживания в БД, видим, что первое вещество можно идентифицировать, как пирен ($V_R(1,1)$=3301 мкл), второе – как ионол ($V_R(1,2)$=3569 мкл), третье – как изомилбензоат ($V_R(1,3)$=3304 мкл). Если с идентификацией второго вещества особых проблем нет, то с идентификацией первого и третьего веществ возникают проблемы, поскольку их можно идентифицировать и как пирен, и как изомилбензоат, при этом ошибка 2-го рода (вероятность пропуска идентификации), вычисляемая по формуле (см., например, [4])

$$\beta = 1 - 2\Phi\left(\frac{d(i,1)}{\sigma(i,1)}\right), \qquad (6)$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$ - хорошо известная функция Лапласа, которая для обоих веществ невелика и принимает значения порядка 0,05 – 0,06. Это означает, что провести однозначную идентификацию веществ по значениям объемов удерживания, перечисленных в (5), не представляется возможным. Это также означает, что для проведения однозначной идентификации необходимо привлекать дополнительную информацию о свойствах веществ, например, спектральные отношения.

Переформулируем задачу идентификации для включения в нее набора спектральных отношений. Необходимо отметить, что идентификацию следует проводить не для каждого спектрального отношения, а для некоторой ин-

тегральной характеристики набора спектральных отношений, как это делается, например, при оценке близости спектров в спектральном анализе [4]. В качестве такой интегральной характеристики предлагается рассматривать среднее значение спектральных отношений $\bar{s}$ конкретного вещества. Заметим, что это среднее значение не является математическим ожиданием, поскольку набор спектральных отношений для каждого вещества не является случайным. Таким образом, среднее значение спектральных отношений $\bar{s}$ вещества является просто средним арифметическим набора значений спектральных отношений.

При решении проблемы идентификации, как и в предыдущем случае, необходимо иметь дело как с базовым набором спектральных отношений, которому будем присваивать индекс 0, так и с экспериментально измеренными наборами (индекс i). Далее, необходимо ввести в рассмотрение разности (невязки $\Delta(0,i,j)$) между базовыми и измеренными значениями спектральных отношений, т.е. для j го вещества вводится следующий набор невязок

$$\Delta(0,i,j) = s(0,i) - s(i,j)\,. \qquad (7)$$

Этот набор невязок уже будет случайным, и, как всякая случайная величина, характеризуется математическим ожиданием, оцениваемым как среднее арифметическое, и СКО. Поскольку объем выборки N невелик (N=7), предполагается, что набор невязок описывается распределением Стьюдента. Введем в рассмотрение среднее значение невязок для j го вещества

$$\Delta_{ср}(0,j) = \frac{1}{N}\sum_{i=1}^{N}\Delta(0,i,j)\,, \qquad (8)$$

где N - число спектральных отношений.

Для согласования вычисленных средних значений невязок со значениями, полученными из данных по средним значениям спектральных отношений, при вычислении средних значений невязок следует брать их алгебраическую сумму.

Из (7) следует, что

$$\Delta_{ср}(0,j) = \bar{s}(0) - \bar{s}(j)\,, \qquad (9)$$

где средние значения наборов спектральных отношений определяются соотношениями

$$\bar{s}(0) = \frac{1}{N}\sum_{i=1}^{N} s(0,i)\,,\ \bar{s}(j) = \frac{1}{N}\sum_{i=1}^{N} s(i,j)\,.$$

Вещество считается идентифицированным, если выполняется условие

$$\Delta_{cp}(0,j) < d\,, \tag{10}$$

или,

$$\bar{s}(0) - \bar{s}(j) < d\,,$$

где d - допуск на среднее значение базовых спектральных отношений: $\bar{s}(0) - d < \bar{s}(0) < \bar{s}(0) + d$.

Если сами базовые спектральные отношения определяли в результате многократных измерений, то допуск на каждое значение спектрального отношения записывается в виде $d(i) = l \cdot \sigma(0)$, где $\sigma(0)$ - среднее квадратическое отклонение, $l = 2$ или 3 (см. формулу (1)) . Допуск на среднее значение спектральных отношений записывают в виде $d = l \cdot \sigma(0)/\sqrt{N}$. Таким образом, условие идентификации по спектральным отношениям принимает вид

$$\Delta_{cp}(0,j) < d = \frac{l \cdot \sigma(0)}{\sqrt{N}}\ . \tag{11}$$

Гипотеза H_0 (нулевая гипотеза) заключается в утверждении, что рассматриваемое вещество в пробе отсутствует. Это означает, что в этом случае должно выполняться неравенство

$$\Delta_{cp}(0,j) - d > 0\,. \tag{12}$$

Гипотеза H_1 (альтернативная гипотеза) предполагает наличие вещества в пробе. Условие присутствия вещества выражается неравенством

$$\Delta_{cp}(0,j) - d < 0\,. \tag{13}$$

Для выяснения вопроса о том, какая из двух гипотез реализуется, предлагается использовать критерий Стьюдента.

Квантиль распределения Стьюдента $t_{N-1,P}$, отвечающий вероятности P = 0,95 и числу степеней свободы N - 1 = 6 (N = 7) равен 2,4470. Это означает,

что именно такое значение переменной $\Delta_{cp}(0,j)-d$ реализуется с вероятностью P = 0,95. Обратим внимание на то, что, если гипотеза H_0 верна, то переменная $\Delta_{cp}(0,j)-d$ принимает положительные значения, при гипотезе H_1 эта переменная отрицательна. Знак этой переменной определяет и знак коэффициентов Стьюдента

$$t(0,j)=\frac{\sqrt{N}\cdot(\Delta_{cp}(0,j)-d)}{S(0,j)}, \tag{14}$$

где S вычисляют по формуле

$$S^2(0,j)=\frac{1}{(N-1)}\sum_{i=1}^{N}(\Delta(0,i,j)-\Delta_{cp}(0,j))^2. \tag{15}$$

Таблицы отрицательных коэффициентов Стьюдента отсутствуют, поскольку такие коэффициенты могут быть получены из положительных коэффициентов с использованием следующего свойства симметрии

$$t_{\alpha}(0,j)=-t_{1-\alpha}(0,j). \tag{16}$$

Если вычисленное по экспериментальным данным значение квантиля $t(0,j)$ оказывается меньше табличного $t_{N-1,P}$, т.е. выполняется условие

$$t(0,j)<t_{N-1,P}\ , \tag{17}$$

то соответствующая гипотеза отвергается, альтернативная – принимается.

Применение критерия Стьюдента поясним на следующем примере.

С этой целью приведем данные по спектральным отношениям для веществ ряда (5).

Пирен: $V_R(1,1)$=3301 мкл, 1.15 3.55 5.77 1.08 1.88 0.40 0.59 $\bar{s}(0)=2{,}06$

Вещество 1: $V(1,1)$ = 3302 мкл, 1.12 3.22 6.00 0.98 1.86 0.50 0.53 $\bar{s}(1)=2{,}03$

Вещество 2: $V(2,1)$ = 3569 мкл, 0.54 0.33 0.06 0.01 0.04 0.15 0.00 $\bar{s}(2)=0{,}16$

Вещество 3: $V(3,1)$ = 3304мкл. 2.15 3.15 1.58 0.27 0.17 0.17 0.00 $\bar{s}(3)=1{,}07$.

Здесь 7 колонок цифр означают спектральные отношения, полученные для анализируемых веществ на восьми длинах волн, правая колонка относится к средним значениям спектральных отношений.

На основании приведенных данных можно по формуле (9) вычислить невязки средних значений спектральных отношений анализируемых веществ по отношению к базовым спектральным отношениям для пирена. После этого по формуле (15) вычисляют СКО невязок для анализируемых веществ и по формуле (14) – значения соответствующих коэффициентов Стьюдента.

Допуск на значения спектральных отношений базового вещества выбирают равным 2σ(0) = 0,02. Поскольку теперь сравнение идет по средним значениям спектральных отношений и невязок, то удвоенное значение СКО, определяющее допуск, будет равно

$$\frac{2\sigma(0)}{\sqrt{N=7}} = d = 0{,}0075$$

В предположении, что реализуется гипотеза H_0 (вещество в пробе отсутствует, и переменная $\Delta_{cp}(0,j) - d$ положительна), для коэффициентов Стьюдента получаем

$$t_0(0{,}1) = 1{,}9875,$$

$$t_0(0{,}2) = 2{,}6395,$$

$$t_0(0{,}3) = 2{,}6292.$$

Напомним, что табличное значение $t_{P=0{,}95} = 2{,}4470$.

Сравнивая вычисленные значения коэффициентов Стьюдента с табличными (см. условие (17)), приходим к выводу, что для веществ 2 и 3 гипотеза H_0 справедлива (эти вещества в пробе отсутствуют), а для вещества 1 эта гипотеза должна быть отвергнута, т.е. оно может быть идентифицировано как пирен. При этом уровень значимости α (значение ошибки 1го рода, т.е. вероятность ложной идентификации) оказывается равным 0,05, а соответствующая вероятность идентификации $P = 0{,}95$.

Аналогичный результат получается и при анализе пробы в рамках гипотезы H_1. В рассматриваемом случае переменная $\Delta_{cp}(0,j) - d < 0$, и соответствующие коэффициенты Стьюдента тоже отрицательны. Табличное значение, отвечающее уровню значимости 0,05, с учетом свойства симметрии коэффициентов Стьюдента (16) также отрицательно, т.е. $t_{P=0{,}05} = -2{,}4470$. Срав-

нивая вычисленные значения коэффициентов Стьюдента с табличным, получаем, что $t_1(0,1) = -1,9875 > t_{P=0,05} = -2,4470$, т.е. вещество 1 удовлетворяет условиям гипотезы H_1. Таким образом, первое вещество может и в этом случае идентифицироваться как пирен. Для остальных двух веществ гипотеза H_1 должна быть отвергнута, т.е. с вероятностью 0,95 их нельзя отождествлять с пиреном.

В заключение следует отметить, что качественно идентификацию веществ с помощью спектральных отношений можно провести по их значениям и распределениям по длинам волн, не прибегая к критерию Стьюдента. Основное достоинство предложенного метода идентификации заключается в возможности количественной оценки ее достоверности.