ЧИСЛЕННЫЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ ДЛЯ ВЫСОКОЭФФЕКТИВНОЙ ЖИДКОСТНОЙ ХРОМАТОГРАФИИ

К.В.Воронцов, Н.В.Кобзева
Информационный консультативный токсикологический центр МЗ РФ
М о с к в а

В статье рассмотрен принципиально новый алгоритм идентификации лекарственных препаратов в биосредах организма при оперативной ректификации и разделении методом жидкостной хроматографии. Алгоритм основан на матричном подходе к обработке числовых данных, поступающих со сканирующего ультрафиолетового (УФ-) детектора. Более полное использование первичных данных в совокупности с математическим моделированием хроматографических пиков позволяет существенно улучшить качество идентификации. Кроме того, становится возможным корректно выделять спектры лекарственных препаратов, отсутствующих в базе данных (БД), хранить в памяти компьютера до десятков тысяч ранее полученных хроматограмм, ускорить самообучение системы, а также повысить чувствительность прибора при предельно низких концентрациях. Все предлагаемые модификации относятся к программному обеспечению и не затрагивают аппаратной части хроматографической установки.

Система экстренной токсикологии REMEDI

Хроматографические методы широко применяются для анализа образцов плазмы крови или мочи, когда требуется исчерпывающая идентификация содержащихся в них лекарственных препаратов. Одной из самых современных технологий в этой области является высокоэффективная жидкостная хроматография в сочетании с оперативной компьютерной обработкой данных. К анализаторам такого класса относится система экстренной токсикологии REMEDI (фирма «Bio-Rad Laboratories», США).

Вкратце принцип действия REMEDI состоит в следующем [6 - 7]. На первой стадии анализа образца происходит промывка системы и удаление из образца белков и других интерферирующих веществ. Вторая стадия, длящаяся 15 минут, представляет собой собственно изократическое разделение. В течение этого времени в компьютер поступают данные от сканирующего УФ-детектора, снимающего на выходе четырехколоночной хроматографической системы спектры в диапазоне 200 - 300 нм. Эти данные, представленные в числовом виде, образуют исходную информацию для третьей стадии — идентификации химических веществ. Алгоритм идентификации сопоставляет спектральные и хроматографические данные образца с базой данных REMEDI, содержащей 257 наименований лекарственных препаратов и 45 метаболитов. На основе этого сопоставления для каждого хроматографического пика

формируется список веществ-кандидатов, из которого методом комбинированного поиска с возвратами выбирается окончательное решение. Более полное описание данного алгоритма можно найти в работе [8].

Опыт эксплуатации REMEDI в Информационно-консультативном токсикологическом центре МЗ РФ, а также анализ зарубежных публикаций позволили сделать следующие выводы относительно программного и математического обеспечения системы.

1. Алгоритм идентификации, сравнивая спектр, взятый из анализируемой хроматограммы, со спектрами из БД, использует в основном «точечные» критерии, такие как совпадение точек максимумов и точек перегиба на спектрах, а также скорректированных отношений [8], вычисленных при некоторых длинах волн. Ясно, что такой подход заведомо ухудшает качество идентификации, т. к. игнорируется значительная часть информации, содержащейся в спектрах (каждый спектр, снимаемый УФ-детектором, состоит из 112 точек). На практике это приводит к тому, что системе не всегда удается идентифицировать вещество, даже если оно присутствует в БД. Такие случаи можно классифицировать следующим образом.

Близкие времена удерживания у двух или большего числа веществ. В результате наложения спектров различных веществ и искажения их первоначальной формы алгоритм с вероятностью около 10% не идентифицирует обнаруженные пики или один из них. Показателен случай, когда не идентифицировался фенобарбетал, т. к. присутствовавшая в нем характерная примесь давала близкий пик.

Последовательность перекрывающихся пиков. При наложении большого числа пиков некоторые из них даже не обнаруживались, а вероятность правильной идентификации заметно снижалась. По этой причине затруднена идентификация многих снотворных и других веществ, элюирующих в начальной части хроматограммы [6].

Низкие концентрации вещества в образце. В этом случае ненадежная идентификация представляется вполне оправданной, т. к. высота пика становится сравнимой с уровнем шума УФ-детектора. Как показали эксперименты [6], при отношении сигнала к шуму, равному 12, правильно идентифицируется не менее 95% пиков. Следует однако учитывать, что в алгоритме не делается попыток сгладить влияние шума путем усреднения спектров, что, как известно из статистики, должно привести к увеличению точности. Таким образом, имеется еще некоторый резерв для снижения порога идентифицируемости.

Вещества с нехарактерным спектром, лишенным точек максимума или точек перегиба. Поскольку эти точки используются при идентификации, их отсутствие приводит к увеличению вероятности неверной идентификации во всех выше перечисленных случаях.

2. Хотя в системе заложена возможность самообучения, в действительности занести в БД хроматографические и спектральные данные неизвестного вещества можно только при наличии отчетливого изолированного пика. На практике это условие

выполняется довольно редко, что сильно снижает эффективность самообучения.

3. В памяти компьютера (на жестком диске) сохраняются результаты предыдущих анализов. Однако если их число превысит максимально допустимое (около 130), все они будут уничтожены и заполнение диска начнется заново. Этот факт препятствует теоретически возможному «накоплению опыта» и связан с тем, что для каждого анализа целиком хранится файл первичных данных размером около 430 кбайт. В то же время практика показала, что часто появляется необходимость просмотреть результаты старых анализов, в том числе и сами исходные данные, например, в виде трехмерного графика.

Причины перечисленных недостатков следует искать только в несовершенстве алгоритма идентификации. Возможным путем их устранения могло бы стать преобразование первичных данных, поступающих с УФ-детектора, с целью выделения истинных спектров и хроматограмм содержащихся в образце веществ. Под истинным спектром понимается тот спектр, который давало бы данное вещество на отчетливом изолированном хроматографическом пике, не подверженном влиянию пиков с близкими временами удерживания. Аналогично определяется понятие истинной хроматограммы.

Предварительная обработка данных позволила бы отказаться от ненадежного алгоритма идентификации, поскольку наличие истинных спектров дает возможность с высокой степенью точности сопоставлять данное вещество с хранящимися в БД. Кроме того, сколько бы неизвестных веществ (отсутствующих в базе данных REMEDI) не содержалось в образце, их истинные спектры все равно были бы вычислены и подготовлены для пополнения БД. Таким образом, появляется реальная возможность построения эффективно самообучающейся системы. Можно проводить и целенаправленное обучение, анализируя заранее подготовленные образцы с известным набором веществ. При этом процесс обучения можно заметно ускорить и удешевить, т. к. в одну пробу можно будет вводить до 10 - 30 различных препаратов. Несколько повысилась бы надежность идентификации веществ с низкой концентрацией, поскольку при вычислении истинных спектров и хроматограмм неизбежно их статистическое усреднение. Наконец, отпала бы необходимость хранить огромные файлы первичных данных в памяти компьютера, т. к. их всегда можно было бы с высокой точностью восстановить по набору спектров и хроматограмм отдельных веществ. При этом для сохранения информации об анализе образца требовалось бы в сотни раз меньших объемов памяти. Таким образом, можно было бы запоминать не 130 предыдущих экспериментов, а всю историю работы системы и использовать накопленный опыт для идентификации.

Рассмотрим один из возможных подходов, позволяющих полностью решить поставленную задачу.

Матричный подход к обработке выходных данных сканирующего УФ-детектора

В основе матричного подхода заложены следующие простые идеи. Необходимо преобразовать файл первичных данных в сжатое описание, представляющее собой набор истинных спектров и хроматограмм всех содержащихся в образце веществ. Математически строго доказывается, что однозначно выполнить такое преобразование можно только при наличии дополнительной информации либо о виде спектров, либо о виде хроматограмм. В данном случае в качестве такой информации используется теоретическое уравнение хроматографического пика с четырьмя свободными параметрами. Это позволяет устранить неоднозначность и вычислить набор спектров и хроматограмм, достаточно близких к истинным.

Название «матричный подход» связано с активным использованием техники теории матриц. Подчеркивается также, что первичные данные обрабатываются единообразно, как большой массив чисел, исключая выделение особых точек или участков данных. Не приводя строгих доказательств, перечислим основные положения матричного подхода.

1. *Структура первичных данных.* Данные, поступающие с УФ-детектора в компьютер, представляют собой числовые значения величины сигналов, регистрируемых детектором в моменты времени $t_1,\dots,t_n$, $n = 1920$, на длинах волн $\lambda_1,\dots,\lambda_m$, $m = 112$, для каждого момента времени. Величина сигнала прямо пропорциональна сумме концентраций проэлюировавших в данный момент веществ, помноженных на некоторые коэффициенты. Эти коэффициенты определяются способностью веществ поглощать излучение определенных длин волн и, очевидно, не зависят от времени. С другой стороны, концентрации веществ в подвижной фазе не могут зависеть от длины волны. Таким образом, первичные данные можно рассматривать как таблицу значений функции

$$z(t,\lambda) = \sum_{s=1}^{k} C_s(t) F_s(\lambda),$$

где k — число химических веществ в образце;

$C_s(t)$ — истинная хроматограмма s-ого вещества;

$F_s(\lambda)$ — истинный спектр s-ого вещества.

Всего в файле первичных данных содержится $nm = 215040$ чисел. Задача состоит в том, чтобы научиться по данной совокупности числовых значений, известных с некоторыми погрешностями (вследствие шума УФ-детектора), определять истинные спектры и хроматограммы для каждого вещества, содержащегося в образце. Заметим, что количество веществ k также неизвестно и подлежит определению.

Запишем полученное соотношение для всех моментов времени при всех длинах волн. Это приведет нас к системе из nm уравнений, которую удобно выразить в виде одного матричного уравнения

$$XY = Z,$$

где $Z = \left\| z(t_i, \lambda_j) \right\|_{i=1,n}^{j=1,m}$ — известная $n \times m$-матрица первичных данных;

$X = \left\| C_s(t_i) \right\|_{i=1,n}^{s=1,k}$ — неизвестная $n \times k$-матрица хроматограмм;

$Y = \left\| F_s(\lambda_j) \right\|_{s=1,k}^{j=1,m}$ — неизвестная $k \times m$-матрица спектров.

Для определения размера k матриц X и Y разумнее всего потребовать, чтобы это было наименьшее число, при котором решения данного уравнения еще существуют. Пару матриц (X, Y), удовлетворяющую всем этим условиям, будем называть разложением исходной матрицы Z.

Элементарное следствие выведенного матричного уравнения: если становится известной матрица хроматограмм, по ней тут же однозначно определяется матрица спектров. И наоборот, по известной матрице спектров можно найти матрицу хроматограмм. Приведем формулы:

$$Y = (X^T X)^{-1} X^T Z, \quad X = ZY^T (YY^T)^{-1}. \qquad (1)$$

2. *Неоднозначность разложения*. Более глубокий анализ задачи разложения показал, что множество ее решений бесконечно. А именно, если Z имеет разложение (X, Y), то всякая пара матриц вида $(XA,\ BY)$ также будет разложением Z, если только произведение квадратных матриц A и B равно единичной матрице. Причем не существует разложений Z, не допускающих представления в таком виде.

Применительно к задаче вычисления истинных спектров и хроматограмм этот результат означает, что на основе одного лишь файла первичных данных однозначно определить их невозможно. Полученное описание множества разложений позволяет лишь выделить классы функций, в которых нам следует их искать. Для того, чтобы из множества разложений выбрать единственное, описывающее истинные спектры и хроматограммы (будем называть это разложение истинным), необходимо привлекать какую-то дополнительную информацию.

Теоретический вопрос, возникший на данном этапе исследований, состоял в следующем. Какого вида дополнительную информацию о спектрах и хроматограммах необходимо использовать, чтобы свести множество разложений к единственному истинному (или очень близкому к истинному) разложению. Было решено, что строить какие-либо гипотезы о спектрах нецелесообразно, т. к. вид этих функций достаточно произволен и может сильно меняться в зависимости от молекулярного строения веществ. Напротив, «внешний вид» хроматографических пиков, определяемый процессом изократического элюирования, до известной степени одинаков.

Рассмотрим математическую постановку вопроса. Пусть имеется семейство функций $G = \{C(t)\}$, построенное на основе тех или иных гипотез относительно вида хроматограмм $C(t)$. Требуется определить, какими свойствами должно обладать это семейство, чтобы для всякой матрицы начальных данных Z в нем можно было найти ровно k функций, вместе описывающих единственное разложение, наиболее близкое к истинному. Это разложение будем

называть оптимальным. Доказано, что для обеспечения единственности оптимального разложения семейство G должно обладать свойствами точности и нелинейности. Не приводя строгих формулировок, поясним смысл этих понятий. Точность семейства функций означает, что любая хроматограмма, которая потенциально может реализоваться в результате элюирования произвольного вещества, должна достаточно точно описываться какой-либо функцией из G. Нелинейность семейства функций означает, что никакая линейная комбинация k функций из G не может снова принадлежать G.

Таким образом, при конкретизации вида хроматограмм $C(t)$ путем задания семейства функций G необходимо следить за выполнением условий точности и нелинейности.

3. *Модельное уравнение пика.* Один из способов построения семейства G состоит в рассмотрении физической модели процесса изократического элюирования и получении на ее основе параметрических уравнений хроматографического пика.

В простейшей классической модели [3] концентрация C адсорбата в подвижной фазе (элюенте) описывается гладкой функцией времени t и расстояния x вдоль направления движения элюента. Функция $C(t,x)$ удовлетворяет дифференциальному уравнению

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} - v(C)\frac{\partial C}{\partial x},$$

учитывающему два основных процесса, протекающих одновременно: диффузию частиц адсорбата в элюенте (D — коэффициент диффузии) и перенос частиц адсорбата со скоростью $v(C)$, изменяющейся в зависимости от концентрации. При некоторых предположениях эта зависимость, определяющая в итоге несимметричность пика, описывается формулой

$$v(C) = \frac{1}{\Gamma(C)\omega_1 + \omega_2},$$

где ω_1, ω_2 — константы, не зависящие от концентрации;

$\Gamma(C)$ — коэффициент Генри.

Известное выражение для коэффициента Генри может быть получено, например, в рамках простейшей теории адсорбции Лангмюра:

$$\Gamma(C) = \frac{b_1}{1 + b_2 C},$$

где b_1, b_2 — коэффициенты, не зависящие от концентрации. Таким образом, зависимость скорости движения частиц адсорбата от их концентрации является дробно-линейной функцией:

$$v(C) = V\frac{C + p}{C + q},$$

где V, p, q — коэффициенты, не зависящие от концентрации.

Приведенное выше дифференциальное уравнение, как показано в [2], может быть решено независимо от конкретного вида функции

v(C). После подстановки в его решение полученной зависимости оно примет вид:

$$C(t) = \frac{k_1}{\sqrt{t}} \exp\left\{ -\frac{1}{k_2 t}\left(k_3 - t\frac{C(t)+p}{C(t)+q} \right)^2 \right\},$$

где k_1, k_2, k_3, p, q — коэффициенты, не зависящие от концентрации и времени. Данная формула не позволяет выразить $C(t)$ в явном аналитическом виде, т. к. является, по существу, трансцендентным уравнением относительно C. Для вычисления $C(t)$ при конкретных значениях параметров и момента времени использовались итерационные численные методы [4].

Показано, что если коэффициент k_1 положить равным 1, то полученное четырех-параметрическое семейство функций G будет удовлетворять введенному условию нелинейности. Менее очевидна точность этого семейства, поскольку при выводе уравнений было сделано немало упрощающих предположений [3], выполнение которых на практике вряд ли возможно. Тем не менее, как показали вычислительные эксперименты по аппроксимации изолированных пиков, в данном семействе всегда находились функции, приближающие пики со среднеквадратичной погрешностью, составляющей не более 2 - 5% от высоты пика.

4. *Алгоритм предварительной обработки*. Изложенные результаты служат обоснованием для следующей схемы обработки первичных данных. Сначала определяется число k и вычисляется одно из разложений, не обязательно оптимальное. Тем самым в нашем распоряжении оказывается, по существу, все множество разложений, т. к. всякое другое может быть получено из данного путем простых линейных преобразований. Затем ищется пересечение полученного множества и множества разложений, у которых хроматограммы описываются выведенной параметрической зависимостью. Точность и нелинейность данного семейства функций гарантируют, что это пересечение состоит из единственного разложения, достаточно близкого к истинному.

Рассмотрим первый этап обработки первичных данных — определение числа веществ в образце с одновременным вычислением одного из разложений (не обязательно оптимального). Простейший итерационный алгоритм может быть построен на основе соотношений (1) между матрицами X и Y. Легко видеть, что, имея некоторую начальную матрицу хроматограмм X, можно подставить ее в первое соотношение и найти матрицу спектров Y, а, зная Y, из второго соотношения получить новое значение для матрицы X. Те же действия можно повторить несколько раз. Оказывается, что при достаточно удачном выборе начального приближения полученный итерационный процесс сходится к одному из разложений Z. Более того, доказано, что погрешность первой итерации прямо пропорциональна, а скорость сходимости обратно пропорциональна среднеквадратичному шуму первичных данных. Поскольку величина шума сканирующего УФ-детектора достаточно мала, описанный

алгоритм сходится, как правило, очень быстро и позволяет достичь высокой точности уже на втором - третьем шаге.

Дальнейшая модификация этого алгоритма позволила автоматически находить приемлемое начальное приближение и определять число k из условия достижения нужной точности разложения. В итоге был получен итерационный численный метод, вычисляющий разложения с точностью, сравнимой с уровнем шума первичных данных и практически не ошибающийся при определении числа веществ в образце. Ввиду громоздкости и относительной сложности алгоритма его подробное описание выходит за рамки данной статьи.

Второй этап предварительной обработки состоит в вычислении оптимального разложения на основе ограничений, накладываемых семейством функций G. Идея решения заключается в следующем. Нам необходимо найти матрицу, столбцы которой представляли бы собой таблицы значений некоторых функций, взятых из семейства G. С другой стороны, в силу линейности множества разложений, каждый из этих столбцов должен образовываться линейными комбинациями столбцов матрицы X из уже вычисленного (не оптимального) разложения. Возникает система уравнений, имеющая, в силу точности и нелинейности семейства G, ровно k решений (в действительности можно говорить лишь о существовании k столбцов, очень близких к решениям, т. к. используемая математическая модель хроматографического пика всегда допускает небольшую погрешность).

Для вычисления столбцов, наиболее близких к решениям, использовался метод наименьших квадратов, который сводился к последовательной минимизации k функций по параметрам k_2, k_3, p, q. Минимум каждой из них находился известными градиентными методами локальной оптимизации [1]. Получавшаяся в результате матрица, составленная из k столбцов, доставляющих минимумы соответствующим функциям, представляла собой матрицу хроматограмм, достаточно близких к истинным. Искомая матрица спектров вычислялась затем по формуле (1).

5. *Идентификация веществ.* По завершении предварительной обработки первичных данных идентификация веществ не представляет особого труда.

Время удерживания для хроматографического пика, заданного модельным уравнением, может быть легко рассчитано по четырем его коэффициентам.

Для сравнения спектров достаточно вычислить их среднеквадратическое или максимальное отклонение. Возможно также построение различных их комбинаций, в том числе параметрических. В последнем случае построение корректного алгоритма идентифи-кации можно осуществить, используя современные методы обучения распознаванию образов [5].

Заметим, что описанный метод предварительной обработки первичных данных позволяет полностью и единообразно использовать все 112 точек сопоставляемых спектров, причем эти спектры достаточно близки к истинным. За счет этого в основном и достигается увеличение качества идентификации веществ.

В целях исследования матричного подхода и тестирования описанных алгоритмов была разработана специализированная программа, позволяющая:

— аппроксимировать изолированные пики элементами рассмотренного параметрического семейства функций;

— вычислять произвольное и оптимальное разложения для заданного участка первичных данных;

— осуществлять преобразование файла первичных данных в более компактный (размером 2 - 8 кбайт) файл-описание оптимального разложения.

— считывать произвольные участки данных из первичного файла или файла-описания и выводить их на экран в виде двумерных и трехмерных графиков.

Программа ориентирована на IBM-совместимый персональный компьютер и реализована на языке программирования Turbo Pascal.

Список литературы

[1] Васильев Ф.П. Численные методы решения экстремальных задач. М.: Наука, 1988. 549 с.

[2] Владимиров В. С. Уравнения математической физики. М.: Наука, 1971. 509 с.

[3] Гольберт К. А., Вигдергауз М. С. Курс газовой хроматографии. М.: Химия, 1974. 407 с.

[4] Демидович Б. П., Марон И. А. Основы вычислитель-ной математики. М.: Физматгиз, 1963. 659 с.

[5] Журавлев Ю. И. // Проблемы кибернетики. 1978. Вып. 33. С. 5-68.

[6] Adams A. K., Essien H., Binder S. R. // Ann. Biol. Clin. 1991. V. 49. P. 291-297.

[7] Binder S. R., Regalia M., Biaggi-McEachern M., Mazhar M. // J. Chromatogr. 1989. V. 473. P. 325.

[8] Fell A. F., Clark B. J., Scott H. P. // J. Chromatogr. 1984. V. 316. P. 423.