ФИЛЬТРАЦИЯ ШУМОВ: ОКОНЧАТЕЛЬНОЕ РЕШЕНИЕ ПРОБЛЕМЫ

Ю.Каламбет, С.Мальцев kalambet@ampersand.ru | ЗАО "Амперсенд"
Ю.Козьмин Институт биоорганической химии РАН

Алгоритмы цифровой обработки сигналов, цифровой фильтрации – важнейший элемент любого измерительного комплекса. В полной мере это относится к современной хроматографии. Авторы предлагают собственный алгоритм цифровой фильтрации, основанный на минимизации доверительного интервала аппроксимации. Метод реализован в программе "МультиХром" (www. multichrom. ru). Приведенные в статье реальные примеры обработки хроматограмм позволяют согласиться с утверждением, что "предложенный алгоритм ставит точку в дискуссии о том, какой метод фильтрации шумов лучше".

Каждое измерение – это сумма полезного сигнала и погрешности (случайной и систематической). Погрешности (шумы) создают электронные системы регистрации, внешние радиопомехи, изменения условий окружающей среды и т.д. Основная задача при обработке цифрового сигнала – получить на основании массива цифровых данных максимально точную оценку полезного аналогового сигнала, породившего эти данные.

Для фильтрации шумов сегодня используется достаточно широкий набор алгоритмов [1]. Ни один из алгоритмов не дает оценку качества сглаживания. Большинство известных методов изменяют форму сигнала. Например, на хроматограмме после фильтрации шумов может измениться форма пиков, и чем лучше результаты фильтрации шумов на базовой линии*, тем более существенно изменение формы пика. Поэтому главная проблема всех методов – нет критерия оценки качества сглаживания.

Нами предложен принципиально новый алгоритм фильтрации шумов (доверительный фильтр), который обеспечивает минимально возможный доверительный интервал для каждой точки массива измерений [2]. Алгоритм ставит точку в дискуссии о том, какой метод фильтрации шумов лучше подходит для каждого конкретного случая, поскольку он самонастраивается и выдает наилучший возможный в каждом случае результат.

Алгоритм фильтрации был опробован на широком спектре реальных анализов разного типа. С его помощью всегда удавалось достичь значительно лучших результатов, чем с помощью метода Савицкого-Голея.

ДОВЕРИТЕЛЬНЫЙ ФИЛЬТР

В основе предложенного алгоритма лежит хорошо знакомый каждому химику-аналитику метод наименьших квадратов. Он применяется, в частности, для построения градуировочных зависимостей. Метод наименьших квадратов основан на аппроксимации последовательности исходных данных линейной комбинацией неких функций, при этом минимизируется сумма квадратов отклонений исходных данных от аппроксимированных значений. Чаще всего в качестве аппроксимирующих функций используются полиномиальные функции (см. врезку). Применительно к дискретной обработке сигнала, аппроксимирующие полиномы в окрестности каждой точки строятся для выборки из исходного массива данных (окно аппроксимации). Ширина этого окна – число точек в выборке.

* Здесь базовая линия – участки хроматограммы, на которых сигнал компонентов, отличных от элюата, не является значимым.

Помимо аппроксимированного значения, метод наименьших квадратов позволяет рассчитать доверительный интервал оценки – насколько далеко аппроксимированное значение отклоняется от "истинного". Мы применили расчет доверительных интервалов к задаче фильтрации шумов.

Среди механизмов фильтрации шумов наиболее популярен метод Савицкого-Голея, основанный на методе наименьших квадратов. Оказалось, что его возможно существенно улучшить. Для каждого аппроксимирующего полинома, построенного вблизи какой-либо точки, можно построить доверительный интервал аппроксимации для этой точки. Среди всех возможных полиномов выбирается тот, доверительный интервал которого в данной точке минимален.

Результатом работы доверительного алгоритма фильтрации шумов является новое "отфильтрованное" значение сигнала и доверительный интервал для каждой из точек массива данных. Безусловно, данный метод будет проигрывать оригинальному методу Савицкого-Голея по вычислительной сложности и скорости выполнения алгоритма, поскольку нужно перебирать некоторое множество полиномов. Однако в условиях избытка вычислительной мощности современных компьютеров это не представляется большой проблемой.

При практической реализации сглаживающего фильтра на основе доверительного алгоритма был выявлен ряд нежелательных эффектов. Во-первых, при малом окне велика вероятность занизить оценку S^2-дисперсии σ^2 (см. врезку), на основе которой вычисляется доверительный интервал. Это происходит, к примеру, если четыре идущие подряд точки случайным образом легли близко к прямой. Такая аппроксимация может быть ошибочно принята за лучшую. Решению проблемы может способствовать априорное знание величины дисперсии ошибки измерения (шума) σ^2. В этом случает в формуле расчета доверительного интервала S заменяется непосредственно на σ.

Кроме того, априорное знание дисперсии шума позволяет проверить правильность аппроксимационной модели. Так, если остаточная сумма квадратов аппроксимации RSS больше допустимой по критерию Пирсона χ^2, модель (аппроксимационный полином) считается неадекватной, это означает, что аппроксимация плохо описывает исходные данные. Поэтому аппроксимирующая функция считается некорректной и отбрасывается.

Чтобы не углубляться в детали, мы не обсуждаем здесь алгоритм оценки параметров шума. Он может

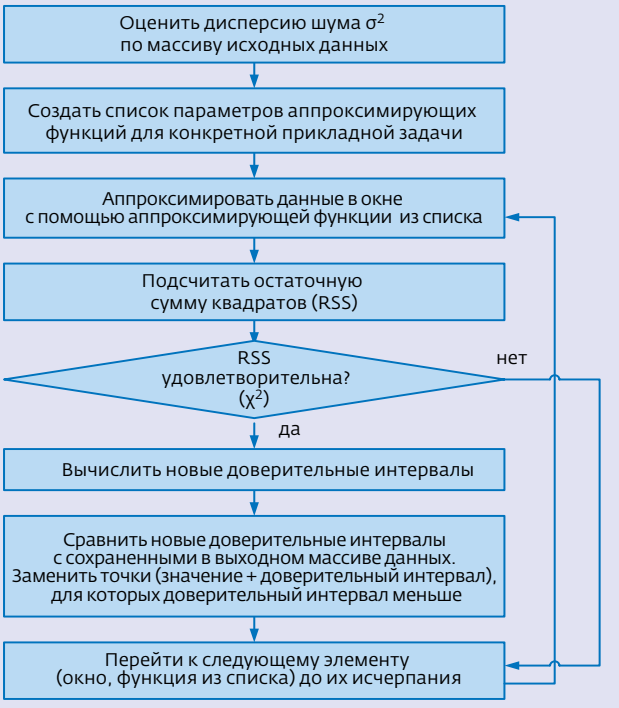
Метод наименьших квадратов

Аппроксимированные значения методом наименьших квадратов [4] вычисляются по формуле $\hat{\mathbf{Y}} = \mathbf{X}\cdot\boldsymbol{\beta}$, где $\hat{\mathbf{Y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}^T$ – вектор аппроксимированных значений,

$$\mathbf{x}=\begin{Vmatrix} 1 & x_1 & x_1^2 & x_1^p \\ 1 & x_2 & x_2^2 & x_2^p \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_3^2 & x_n^p \end{Vmatrix}$$ – матрица степеней по оси x (временной)

$\boldsymbol{\beta} = (X^T\cdot X)^{-1}\cdot X^T\cdot\mathbf{Y}$ – вектор коэффициентов;
$\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}^T$ – вектор откликов детектора;
n – число точек, использованных при построении полиномиальной аппроксимации (ширина окна фильтра);
p – степень полинома + 1.
Доверительный интервал Δ_Y вычисляется как $\Delta_Y = t_{n-p}^{(1/2)\alpha}\cdot S\cdot\sqrt{u_*}$;
$S^2 = RSS / (n-p)$ – оценка дисперсии σ^2 ошибки данных;
$RSS = (\mathbf{Y}-\mathbf{X}\cdot\boldsymbol{\beta})^T\cdot(\mathbf{Y}-\mathbf{X}\cdot\boldsymbol{\beta})$ остаточная сумма квадратов (Residual Sum of Squares)
$u_* = \mathbf{x}_*^T(\mathbf{X}^T\cdot\mathbf{X})^{-1}\mathbf{x}_*$;
$\mathbf{x}_*^T = \{1, x_*, \ldots, x_*^p\}$;
x_* – положение, в котором производится аппроксимационная оценка;
t_m^δ – коэффициент Стьюдента для доверительной вероятности $(1-\delta)$ и m степеней свободы.
При априорно известной дисперсии формула доверительного интервала принимает вид $\Delta_Y = t_{n-p}^{(1/2)\alpha}\cdot\sigma\cdot\sqrt{u_*}$.

Алгоритм работы доверительного фильтра



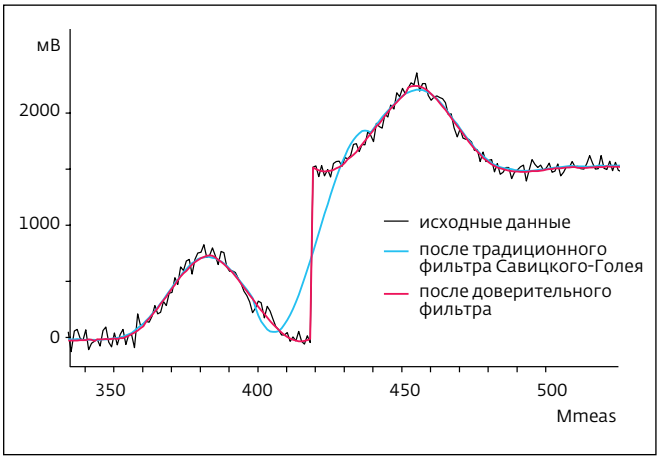


Рис.1. Фильтрация скачка базовой линии с помощью традиционного фильтра Савицкого-Голея и доверительного фильтра. В обоих случаях ширина окна равна 41

быть аналогичен описанному нами ранее [5] или каким-либо иным, но его главная задача – дать оценку величины дисперсии шума. При этом принимается предположение, что дисперсия шума σ^2 одинакова по всему массиву анализируемых данных и оценивается с использованием всего массива данных.

РЕАЛИЗАЦИЯ ДОВЕРИТЕЛЬНОГО АЛГОРИТМА ФИЛЬТРАЦИИ В ПРОГРАММЕ "МультиХром"

Доверительный фильтр реализован в программе "МультиХром". В ней имеется настроечный параметр (Noise Definition Window, NDW), предназначенный для определения дисперсии шума и выглядящий для пользователя как размер окна фильтрации. Этот размер используется для оценки дисперсии шума по всей хроматограмме, его рекомендованная величина соответствует полуширине типового склона пика. При аппроксимации хроматограммы используются кубические полиномы (p = 4), величина окна изменяется от NDW/ $(2\sqrt{2})$ до NDW·$2\sqrt{2}$, на каждом шаге увеличиваясь в $\sqrt{2}$ раз, т.е. размер окна изменяется восьмикратно.

Вычислительная сложность алгоритма оказалась невысока, фильтрация шумов типовой 20-минутной хроматограммы, измеренной с частотой опроса 10 Гц, происходит почти незаметно для пользователя даже при использовании процессора Intel Atom.

Рассмотрим некоторые примеры, демонстрирующие эффективность реализации фильтра.

Аппроксимация скачка базовой линии. Большинство существующих фильтров справляется с такой ситуацией плохо или очень плохо. Скачок в большинстве методов фильтрации шумов оказывает влияние на близлежащие точки; в случае доверительного фильтра такого "дальнодействия" нет (рис.1).

Белый шум. Доверительный фильтр демонстрирует оптимальное подавление шума базовой линии без изменения формы пика. Рассмотрим хроматограмму в виде экспоненциально-модифицированной Гауссианы [7] с наложенным белым шумом (рис.2). Минимальная ширина щели, соответствующая середине окна, равна 5, максимальная – 40. Если бы мы использовали метод Савицкого-Голея, то для того чтобы избежать деформации формы пика, пришлось бы применить минимальную ширину окна, т.е. 5. Число параметров p кубического полинома равно 4. Дополнительное шумоподавление на базовой линии – $[(40-4)/(5-4)]^{1/2}$ = 6. Ошибка измерения высоты пика уменьшается в 1,4 раза за счет более точного проведения базовой линии.

Пульсации насоса. Пульсация насоса может быть как сигналом, так и шумом, в зависимости от постановки задачи. Если нас интересуют только хроматографические пики, то при высоком заданном априорном уровне шумов пульсации насоса исчезают (рис. 3, кривая б). Если же интересует форма пульсации базовой линии, вызванной работой насоса, то при

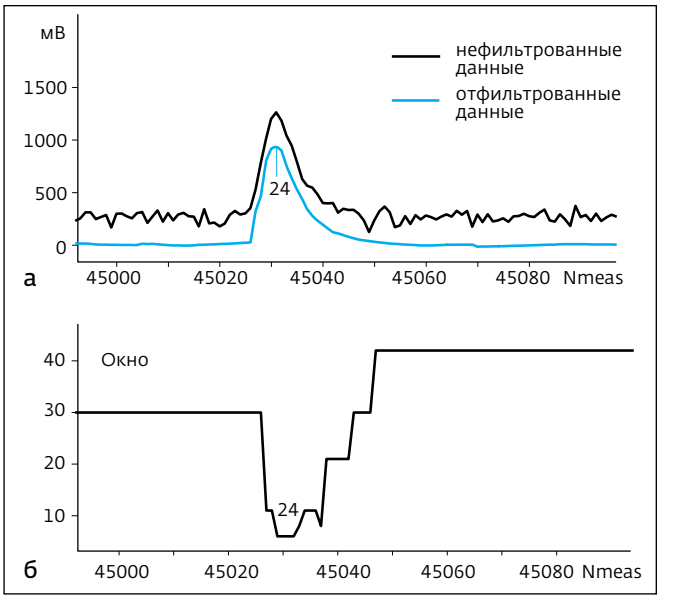


Рис.2. Хроматограмма в виде экспоненциально-модифицированной Гауссианы с наложенным белым шумом: а – нефильтрованные и отфильтрованные данные; б – профиль изменения окна аппроксимационного полинома

низком априорном уровне шумов пульсации остаются, но становятся более гладкими (рис.3, кривая в). При подавлении пульсаций насоса шум не является белым, но несмотря на это использование доверительного фильтра дает приемлемые результаты. Следует иметь в виду, что доверительный интервал сглаженной кривой будет достаточно большим. Гораздо лучших результатов можно достичь, если составить модель процесса и удалить пульсации насоса перед применением доверительного фильтра.

Капиллярный электрофорез – один из самых сложных объектов для применения фильтров из-за формы пиков: они бывают очень узкими или треугольными. Доверительный фильтр хорошо справляется с электрофореграммами. Оптимальное увеличение чувствительности и предела обнаружения не ведет к искажению больших пиков. В примере на рис.4. фильтр Савицкого-Голея с любой шириной окна дает неудовлетворительные результаты (рис.4а). Шумоподавление, обеспеченное доверительным фильтром на базовой линии, будет близко к $(149–4)^{1/2} \approx 12$.

ПРЕДЕЛЫ ОБНАРУЖЕНИЯ И ОПРЕДЕЛЕНИЯ

Согласно официальным документам [6], в хроматографии пределы обнаружения и определения рассчитываются сравнением шума базовой линии с высотой пика. Такая методология не учитывает оценку ошибки измерения в каждой точке и, возможно, подлежит пересмотру. Однако если руководствоваться этой методологией, то применение доверительного фильтра способно увеличить отношение сигнал/шум по сравнению с оптимально подобранным фильтром Савицкого-Голея, приблизительно в $\sqrt{(n_{baseline} - p)/(n_{peak} - p)}$ раз, где $n_{baseline}$ – ширина окна аппроксимации на базовой линии, n_{peak} – минимальная ширина окна аппроксимации в пределах пика. Этот показатель в реальности может составлять 1,5–6 раз. Коэффициент зависит от наличия других пиков вблизи анализируемого, ширины пика и характера шума. Наличие на хроматограмме пиков разной ширины может заметно поднять этот коэффициент, поскольку доверительный фильтр обеспечивает оптимальную аппроксимацию каждого пика, а настройки фильтра Савицкого-Голея при-

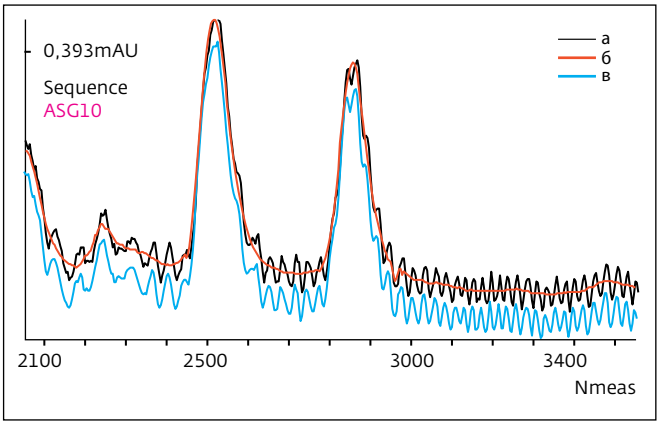


Рис.3. ВЭЖХ-хроматограмма с заметным уровнем помехи, вызванной работой насоса (а), хроматограмма, обработанная доверительным фильтром с большим уровнем шумов (параметр NDW соответствует нескольким циклам насоса) (б) и (в) отфильтрованная доверительным фильтром с низким уровнем шумов (параметр NDW соответствует половине цикла насоса)

шлось бы подбирать для самого узкого из пиков хроматограммы.

Зная доверительные интервалы всех точек хроматограммы, не составляет труда подсчитать ожидаемый доверительный интервал оценки высоты

$$\Delta_{height} = \sqrt{\Delta^2_{\hat{Y}(x_h)} + \Delta^2_{baseline}}$$

и площади пика

$$\Delta_{area} = \sqrt{\sum_{peak} \Delta^2_i + \sum_{peak} \Delta^2_{baseline}}\ .$$

Здесь $\hat{Y}(x_h)$ – аппроксимированное значение хроматограммы в положении, соответствующем вершине пика. Суммирование квадратов доверительных интервалов Δ_i^2 проводится по всем точкам хроматограммы, относящимся к области пика. Возможно, именно эти формулы следует использовать для определения пределов обнаружения и определения.

ПРОБЛЕМЫ И ИТОГИ

Существующая реализация доверительного фильтра не производит отбраковку выбросов – она, скорее, избегает их, оставляя "выскочившие" точки нетронутыми. При необходимости несложно модифицировать процедуру, основываясь на методах устойчивой регрессии [8]. Однако существует опасность, что при таком подходе устойчивая регрессия может скрыть ошибки выбора неправильной модели аппроксимации. Поэтому пока мы предпочитаем не применять алгоритм доверительного фильтра для фильтрации выбросов, используя для

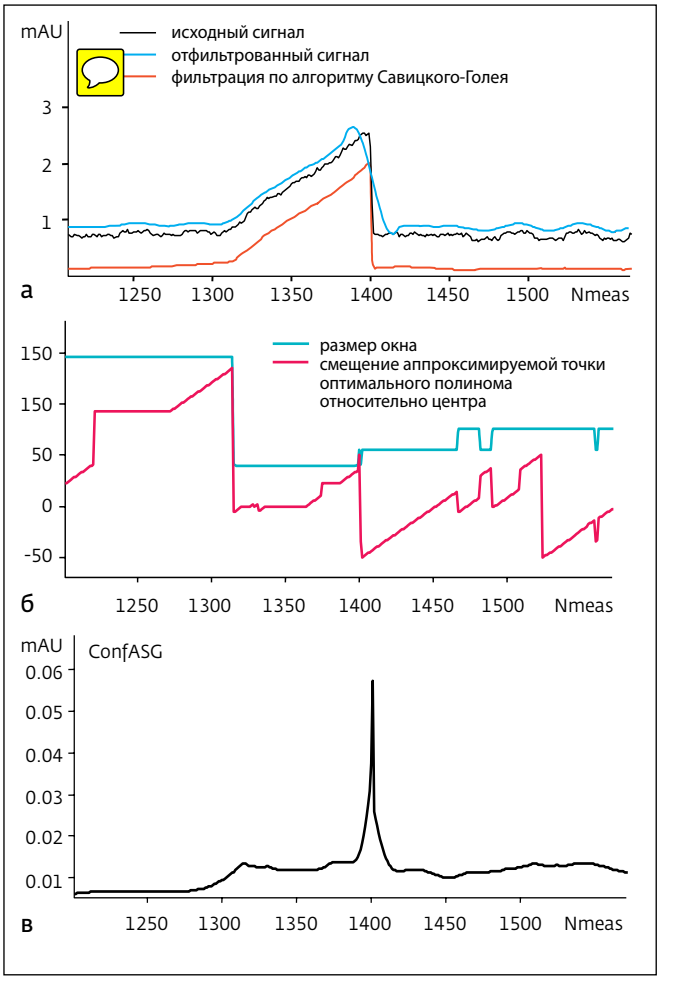


Рис.4. Электрофореграмма: а – исходный сигнал, отфильтрованный сигнал и фильтрация по алгоритму Савицкого-Голея; б – размер окна и смещение аппроксимируемой точки оптимального полинома относительно центра; в – профиль величины доверительного интервала

этого методы, зависящие от контекста. Так, одним из вариантов фильтрации выбросов может быть исключение некоторых точек с аномально большим доверительным интервалом из набора данных, используемых при аппроксимации.

Доверительный интервал – это вполне естественный критерий качества аппроксимации, он прекрасно соответствует задаче фильтрации шумов. В случае переменной ширины окна или степени полинома необходима априорная оценка шума, чтобы избежать эффектов случайной хорошей аппроксимации для малых окон и подавления пиков в случае больших окон. Алгоритм доверительного фильтра очень эффективно подавляет шумы базовой линии и значительно улучшает пределы детектирования и определения. Могут подавляться даже шумы, отличные от белого – такие как пульсации насосов, хими-

ческий шум. Форма пиков при этом не страдает. Безусловно, методы, которые используют дополнительную информацию о сигнале и шуме, такие как восстановление профиля пульсации насоса и вычитание этого профиля из сигнала перед фильтрацией шумов, могут дать лучший результат, но они будут гораздо более трудоемкими, поскольку потребуют конструирования модели процесса.

Математические методы, использованные для вычисления доверительного интервала, базируются на предположении равномерного (гомоскедастичного) шума. Случай неравномерного (гетероскедастичного) шума гораздо более сложен [8, 9] и в некоторых ситуациях может быть сведен к случаю гомоскедастичного шума путем масштабирования сырых данных. В качестве примера процедуры такого масштабирования можно предложить извлечение квадратного корня из величины сигнала перед сглаживанием, с последующим возведением в квадрат результата фильтрации. В случае любого счетного детектора такая процедура имеет хорошее теоретическое обоснование. Пока теории построения доверительного интервала при произвольном гетероскедастичном шуме нам не встречалось.

Доверительный фильтр – лучшая с метрологической точки зрения аппроксимация результатов измерения полиномами. В перспективе аналогичный подход может быть применен к анализу многомерных данных, таких как отклик масс-спектрометрического детектора или фотография. Большие перспективы у использования доверительного интервала в анализе данных. Можно применять аппроксимации, отличающиеся от полиномиальных. В целом методика открывает широкое поле деятельности в будущем.

ЛИТЕРАТУРА

1. **Felinger A.** Data analysis and signal processing in chromatography / Data Handling in Science and Technology, v.21. – Elsevier, 1998.
2. PCT patent publication WO 2011/106527.
3. **Savitzky, A.; Golay, M.J.E.** Smoothing and Differentiation of Data by Simplified Least Squares Procedures. – Analytical Chemistry, 1964, v.36(8), p.1627–1639.
4. **George A. F. Seber and Alan J. Lee.** Linear Regression Analysis. – Wiley, 2003.
5. **Каламбет Ю.А., Михайлова К.В.** Оценка величины шума и ее использование при обработке хроматографического сигнала. – Лабораторный журнал, №1 (1), 2002, с.32–35.
6. The European Pharmacopoeia, 6th ed. – Council of Europe European (COE), European Directorate for the Quality of Medicine, 2007, www.edqm.eu.
7. **McWilliam, I. G.; Bolton, H. C.** Instrumental Peak Distortion. I. Relaxation Time Effects. – Analytical Chemistry, 1969, v.41, p.1755–1762.
8. **Ricardo Maronna, Doug Martin and Victor Yohai.** Robust Statistics - Theory and Methods. – Wiley, 2006.
9. **Schwartz L.M.** Calibration curves with non-uniform variance. – Analytical Chemistry, 1979, v.51(6), p.723–727.