

МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ В ПРИБОРОСТРОЕНИИ

УДК 543.07

EDN: PDFNMQ

© Б. В. Бардин, А. Я. Логинов, В. В. Манойлов, А. И. Петров, 2025

МЕТОД ОБРАБОТКИ СИГНАЛОВ НЕСТАЦИОНАРНЫХ ПРОЦЕССОВ СПЕКТРОМЕТРИИ В ЗАДАЧЕ СЕКВЕНИРОВАНИЯ ДНК С ИСПОЛЬЗОВАНИЕМ ОБРАТНЫХ ФУНКЦИЙ

Необходимость анализа нестационарных процессов возникает во многих задачах спектрометрии. Нестационарными являются процессы, в которых некоторые параметры сигналов меняются во времени. В большом количестве разнообразных приборов время может пониматься условно. Например, в спектрофотометрии это длина волны фотона, в масс-спектрометрии — молекулярная масса, в фотоэлектронной спектрометрии — энергия связи электрона в атоме и т.п. В качестве меры времени в большинстве приборов может быть выбрана физическая характеристика, в которой калибруется шкала спектра — ось абсцисс графика. В настоящей работе рассматриваются алгоритмы обработки сигналов нестационарного процесса, которым является совокупность сигналов сиквенс-анализа ДНК по Сэнгеру методом капиллярного электрофореза. В идеальных случаях в таких приборах ширина спектральных пиков и интервал времени между выходом пиков соседних (в цепи ДНК) нуклеотидов одинаковы для всех каналов. Однако в реальных приборах при анализе длинных фрагментов ДНК порядка 1000 нуклеотидов значения этих параметров изменяются во времени. Так, например, в конце рабочего цикла анализа ширина пиков увеличивается, а расстояние между ними уменьшается, и спектральные пики практически сливаются друг с другом. В работе рассматриваются алгоритмы преобразования нестационарного спектра в стационарный путем переноса точек (отсчетов) исходного нестационарного спектра на шкалу выходного стационарного спектра при помощи обратной функции шкалы времени. Обратная функция позволяет скорректировать нелинейную шкалу времени. Алгоритмы протестированы на математических моделях и на реальных сигналах. Результаты тестирования представлены в работе.

Ключевые слова: секвенирование нуклеиновых кислот, математическая обработка нестационарных процессов в спектрометрии, использование обратных функций

ВВЕДЕНИЕ

Задача обработки сигналов нестационарных процессов возникает во многих научных и технических исследованиях. В спектрометрии основная, "полезная" часть сигнала, как правило, представляет последовательность во времени неотрицательных спектральных пиков. Остальные составляющие сигнала, в частности фоновая составляющая, шумы, помехи, обычно удаляются на начальных этапах обработки. При этом на последующих этапах часто возникает необходимость анализа нестационарных процессов. Это такие процессы, в которых какие-либо параметры сигнала меняются во времени. Причем время здесь может пониматься условно. В общем случае это та физическая характеристика, в которой калибруется шкала спектра — ось абсцисс графика. Так, в фото-спектрометрии это длина волны фотона, в масс-спектрометрии — молекулярная масса, в фотоэлектронной спектрометрии — энергия связи электрона в атоме и т.п.

Наиболее характерным примером таких спектральных сигналов является совокупность сигналов сиквенс-анализа фрагментов ДНК по Сэнгеру методом капиллярного электрофореза [1, 2]. Фореграмма — зарегистрированные в результате электрофореза данные, состоит из четырех спектров, или четырех цветовых каналов, в единой шкале времени. Основные параметры сигналов — ширина спектральных пиков и интервал времени между выходом пиков соседних (в цепи ДНК) нуклеотидов — одинаковы для всех каналов. Однако при анализе длинных фрагментов ДНК, порядка 1000 нуклеотидов, значения параметров существенно изменяются во времени. Так, в конце рабочего цикла анализа ширина пиков увеличивается, а расстояние между ними уменьшается, и спектральные пики практически сливаются друг с другом.

Параметры большинства операций обработки спектров (фильтрации и др.) должны быть согласованы с параметрами полезных сигналов и шумов. Так, например, в выражение описания

обостряющего фильтра [3–5], используемого для разделения слившихся пиков, включается образ искомого полезного пика, параметрами которого являются ширина пика и параметры или математическое выражение формы пика (например, гауссиан). Параметры образа напрямую определяют точность и корректность разделения слившихся пиков сигнала.

Корректность можно проиллюстрировать следующим образом. Если мы анализируем пачку из двух одинаковых слившихся пиков, то при равенстве ширины пика сигнала и образа в фильтре, как правило, выделяются два пика (в случае, если мы работаем в пределах разрешающей способности метода). Если же используется образ в фильтре шириной, в два раза меньшей ширины пика сигнала в данной точке спектра (что вполне может быть), то обычно получаются три пика.

Для достоверной оценки параметров обрабатываемых спектральных сигналов нестационарных случайных процессов требуется найти и использовать такой метод коррекции спектров, который позволил бы оценивать параметры спектров с такой же точностью, как в условиях стационарных процессов. В работе предлагается один из возможных методов оценки параметров спектральных сигналов нестационарных процессов, который основан на использовании обратных функций [6], [7, pp. 103–120]. Метод апробирован в задачах обработки сигналов в программном обеспечении отечественного секвенатора ДНК.

ПРЕДЛАГАЕМОЕ РЕШЕНИЕ

Преобразование нестационарного спектра в стационарный

Такое преобразование при некоторых условиях может быть выполнено переносом точек (отсчетов) исходного нестационарного спектра на шкалу выходного стационарного спектра со сдвигом по времени, пропорциональным отклонению интервала параметра нестационарности от некоторого значения, заданного как соответствующий постоянный интервал выходного спектра. Этот последний интервал в частном случае может быть равным среднему значению интервала в исходном спектре (тогда длительность полученного спектра будет равна длительности исходного) или же выбран из некоторых других соображений.

Преобразование шкалы времени

Задача решается при следующих условиях:

– *Параметр нестационарности*, который в результате преобразования требуется сделать постоянным, представляет собой положительный, нену-

левой интервал времени на всей шкале. Например, такой параметр, как ширина спектрального пика, не может быть отрицательным или нулевым — это не имеет физического смысла.

– *Шкалирующая функция*, описывающая изменение параметра по шкале, — медленная и гладкая (см. далее). Например, для спектра в десятки тысяч отсчетов такая функция вполне приемлемо может быть представлена полиномом 5-го порядка, при том что ширина спектральных пиков (параметр нестационарности) обычно имеет порядок 5–30 отсчетов.

Цифровые оценки параметров приведены применительно к задаче капиллярного электрофореза для анализа ДНК.

Ниже использованы следующие обозначения:

t — "физическое" время, определяет шкалу исходного, зарегистрированного спектра;*)

τ — "виртуальное" время, определяет шкалу "виртуального", идеального спектра, где интервал параметра нестационарности постоянный (constant);

δ — параметр нестационарности на шкале t , например, ширина спектрального пика, интервал времени;

Δ — значение параметра нестационарности на "идеальной", виртуальной шкале τ , постоянный (constant) интервал времени.

Представим себе следующую мысленную модель системы, в которой необходимо решить обсуждаемую задачу.

Входной процесс представляет собой "идеальный" спектр с постоянной шириной спектральных пиков Δ в течение всего времени τ . Однако при его регистрации прибором возникает искажение шкалы времени t выходного зарегистрированного спектра, например, за счет "плавания" частоты задающего генератора, формирующего отсчеты оцифровки выходного спектра. Соответственно меняется в количестве отсчетов ширина спектральных пиков $\delta(t)$ на этой шкале.

В качестве оправдания выбора такого примера скажем, что он наиболее наглядный, хотя и наименее вероятный, так как задающие генераторы являются, пожалуй, самыми стабильными и точными электронными компонентами.

Искажение шкалы описывается некоторой шкалирующей функцией $t = f(\tau)$ (см. рис. 1).

*) Здесь использовано такое имя, чтобы не путать, например, с таким устоявшимся термином, как "реальное" время (on-line). Следует отметить, что обработка спектрометрической информации, как правило, производится как раз в "нереальном" времени (off-line), когда весь спектр уже полностью накоплен и зарегистрирован.

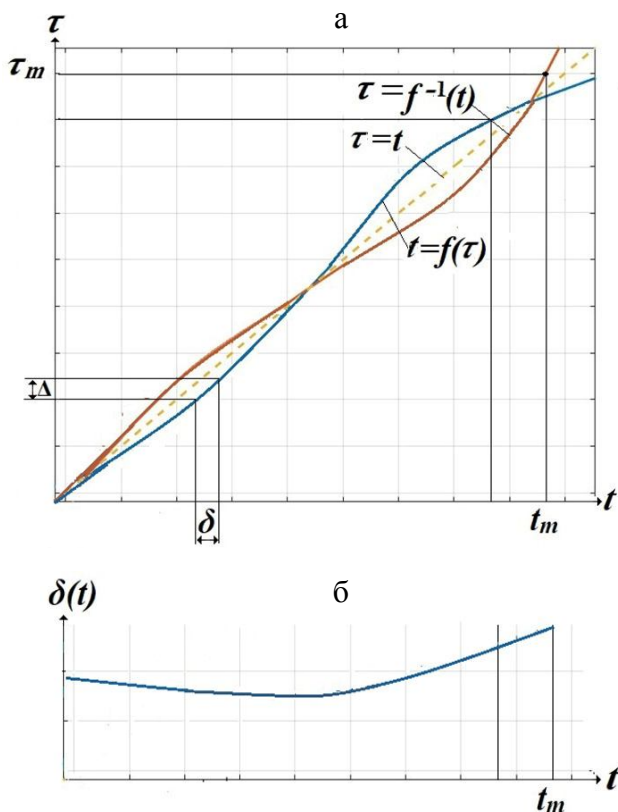


Рис. 1. Пояснения к построению шкалы с помощью обратной функции.
а — шкалирующая и обратная функции,
б — зависимость ширины пиков от времени

Ставится задача восстановления исходного идеального спектра по зарегистрированному спектру на шкале t .

Если функция $f(\tau)$ строго монотонная, то для нее существует обратная функция $\tau = f^{-1}(t)$, которая также является строго монотонной. А из теории обратных функций [6], [7, pp. 103–120] известно, что

$$f^{-1}(f(\tau)) = \tau, \quad f(f^{-1}(t)) = t. \quad (1)$$

Таким образом, для того чтобы в нашей модели восстановить, а в реальности создать, "идеальную" шкалу τ , необходимо подвергнуть физическую шкалу t обратному преобразованию функцией $f^{-1}(t)$.

В реальности в нашем распоряжении имеется зарегистрированный физический спектр на шкале t . Этот спектр можно обработать, найти спектральные пики и построить по ним зависимость ширины пиков от времени $\delta(t)$ (рис. 1).

На рис. 1 величины Δ и δ изображены условно-иллюстративно, чтобы можно было что-то разгля-

деть, реально же эти величины на 2–3 десятичных порядка меньше длины шкалы. Поэтому их вполне приемлемо использовать как дифференциалы. Тогда из рис. 1 производная τ по t будет

$$\frac{d\tau(t)}{dt} = \frac{\Delta}{\delta(t)}. \quad (2)$$

Функция $f^{-1}(t)$ рассчитывается интегрированием ее производной

$$f^{-1}(t) = \tau(t) = \Delta \int_0^t \frac{1}{\delta(t)} dt. \quad (3)$$

Функция $1/\delta(t)$ является ненулевой положительной функцией, и интеграл от нее будет строго монотонным. Поэтому выполняется приведенное выше условие, чтобы функция $f^{-1}(t)$ была строго монотонной.

Функцию $f(\tau)$ можно вычислить подобным же способом. Однако наиболее простым и надежным (в плане устранения возможного накопления ошибок вычислений при интегрировании) способом является формирование f как зеркальное отражение (или результат вращения на 180 градусов) функции f^{-1} относительно оси $\tau = t$. Дело в том, что требование к точности вычисления этих функций, вообще говоря, вполне умеренное, но точность их соотношения, точность выполнения преобразований (1) должна быть абсолютной.

После создания шкалы τ в нее переносятся спектральные данные из шкалы t . Для этого на шкале τ задаются отсчеты τ_i , как правило, с постоянным шагом. Для каждого из этих отсчетов при помощи функции $f^{-1}(t)$ вычисляется соответствующее ему время t_j и соответствующее этому времени значение спектральных данных $D[t(t_j)]$ переносится в отсчет i шкалы τ — $D[\tau(\tau_i)]$. Необходимо отметить, что время t_j почти всегда не будет точно совпадать с временем какого-либо отсчета на шкале t . Поэтому значение $D[t(t_j)]$ необходимо вычислять при помощи интерполяции между соответствующими соседними отсчетами шкалы t .

Пример преобразования шкалы времени для модели спектра фореграммы, содержащей пики разной ширины

Рассмотренный в предыдущем подразделе способ преобразования временной шкалы был сначала проверен с помощью модели фореграммы из 25 000 отсчетов, в которой содержалось 160 пиков гауссовой формы. Пики находились на одинаковом расстоянии, но каждый из них отличался шириной. Изменение ширины пиков от времени происходило по линейному закону.

Следует обратить внимание на то, что в дискретной математике и в программировании

элементы массивов данных идентифицируются номерами этих элементов в массиве. Поэтому для простоты на приведенных ниже рисунках шкалы t и τ оцифрованы в отсчетах и так же на рисунках названы. Шкалы ширины пиков оцифрованы в количестве отсчетов. Секундам эти оцифровки будут соответствовать при равенстве расстояния между отсчетами 1 с. Однако экспериментальные данные, использованные в данной работе, были получены на секвенаторе с заданием интервала между отсчетами, равным 0.25 с. То есть в секундах величина этих оцифровок будет в 4 раза меньше.

На рис. 2, а, показан фрагмент модели спектра, на котором обозначена ширина пиков для начала и конца шкалы до применения алгоритма линеаризации шкалы. На рис. 2, б, приведена зависимость

ширины пиков от времени для модели спектра, представленной на рис. 2, а. На рис. 2, в, — фрагмент модели спектра, на котором показана ширина пиков для начала и конца шкалы после восстановления шкалы времени по рассмотренному выше способу, использующему обратную функцию. На рис. 2, г, показаны зависимости ширины пиков для исходной нелинейной шкалы и для шкалы, восстановленной с помощью обратной функции. При восстановлении шкалы времени использовалась интерполяция на основе полиномов 2-го порядка.

Как видно из рис. 2, г, зависимость ширины пиков от времени в диапазоне от 0 до 20 000 представляет собой линию, близкую к прямой, параллельной горизонтальной оси.

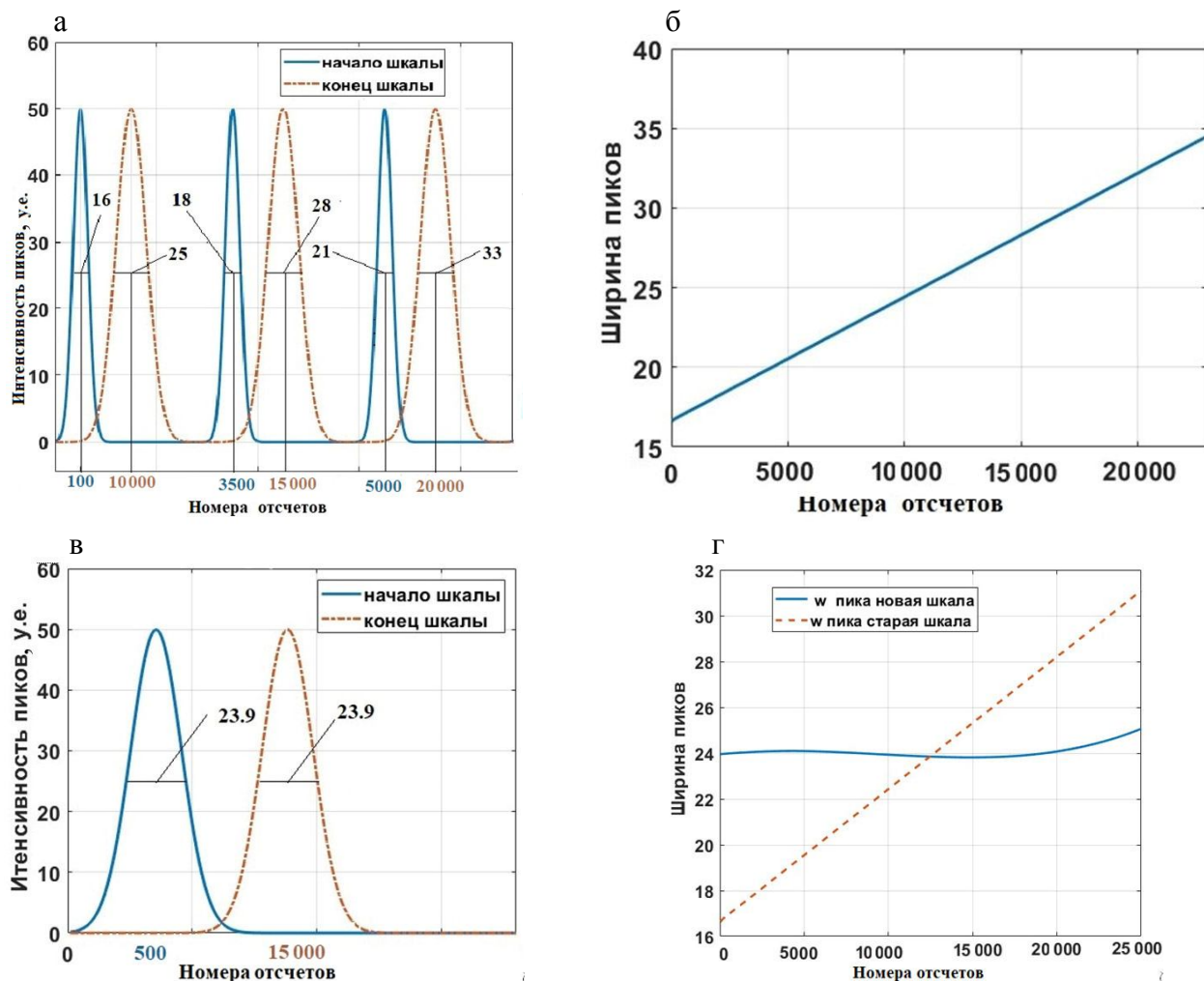


Рис. 2. Фрагменты модели фореграммы и зависимости ширины пиков.

а — фрагмент модели фореграммы с пиками разной ширины в начале и в конце шкалы, ширина пиков дана в количестве отсчетов; б — зависимость ширины пиков от номеров отсчетов; в — фрагмент модели спектра с указанием ширины пиков для начала и конца шкалы после восстановления ее линейности; г — зависимости ширины пиков w для исходной нелинейной шкалы и для шкалы, восстановленной с помощью обратной функции

Среднее значение ширины пиков на этом участке равно 23.9, а среднее квадратичное отклонение — 0.1. На участке от 20 000 до 25 000 можно заметить небольшое увеличение ширины пиков, связанное с краевыми эффектами. Среднее значение ширины пиков во всем диапазоне модели спектра равно 24.0, а среднее квадратичное отклонение — 0.4.

Проведенное на модели исследование способа восстановления линейности шкалы с помощью обратной функции показало его работоспособность для обработки реальных фореграмм.

Пример линеаризации шкалы времени для фореграммы, полученной на секвенаторе ДНК

На рис. 3, а, б, представлены фрагменты фореграммы, полученные на секвенаторе "Нано-

фор-05": а — фрагмент в начале шкалы, б — фрагмент в конце шкалы. В конце шкалы ширина пиков увеличилась.

На рис. 4 представлена зависимость ширины пиков в количестве отсчетов на полувысоте пика. Сплошной линией показана аппроксимация этой зависимости полиномом 5-й степени. Аппроксимация этой зависимости была выполнена методом наименьших квадратов для построения линии регрессии. Параметры функции аппроксимации являются характеристикой нелинейности временной шкалы

На рис. 5 представлены прямая и обратная функции шкалы времени фореграммы. Прямая функция является результатом интегрирования функции аппроксимации ширины пиков.

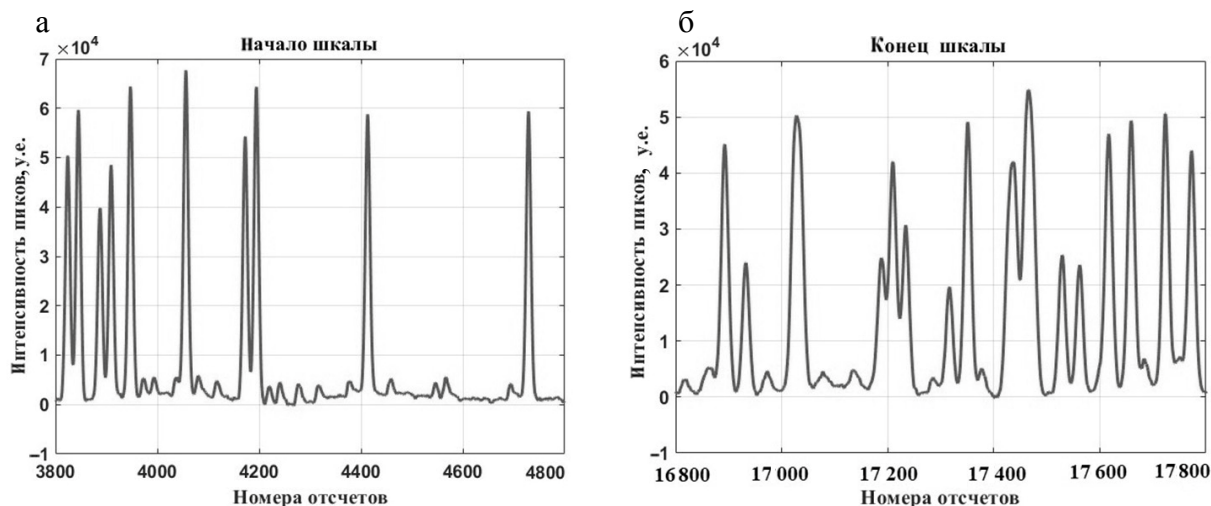


Рис. 3. Фрагменты фореграммы: а — в начале шкалы, б — в конце шкалы

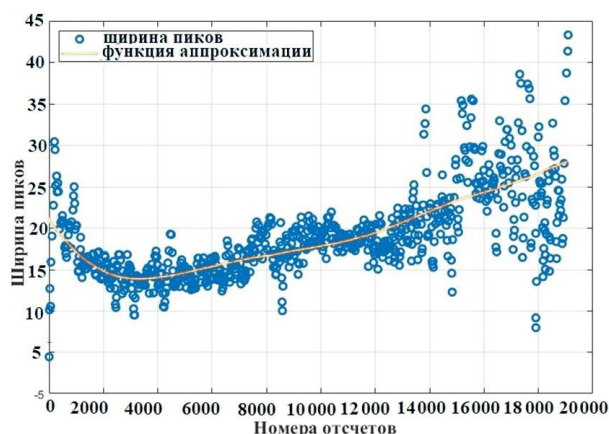


Рис. 4. Зависимость ширины пиков от номеров отсчета.
Сплошная линия — аппроксимация полиномом

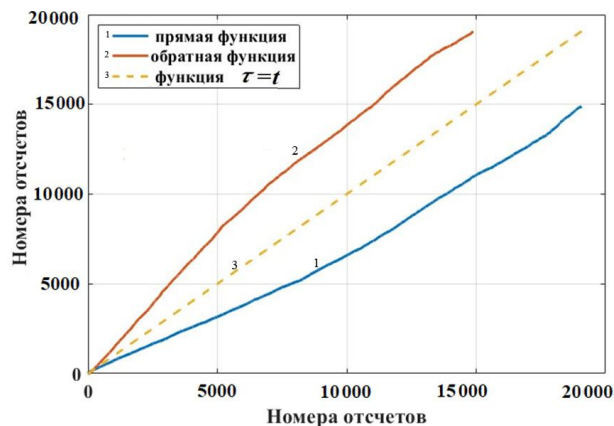


Рис. 5. Прямая $t = f(\tau)$ и обратная $\tau = f^{-1}(t)$ функции шкалы времени фореграммы

Интеграл от функции аппроксимации времени между пиками характеризует шкалу времени и является $t = f(\tau)$. Следуя алгоритму, рассмотренному в предыдущем разделе, найдем обратную функцию $\tau = f^{-1}(t)$, заменяя зависимость переменной по вертикальной оси от переменной по горизонтальной оси зависимостью переменной по горизонтальной оси от переменной по вертикальной оси.

Вычисление значения $D[t(t_i)]$ производилось описанным выше способом с использованием для интерполяции $D[t(t)]$ между соседними отсчетами полинома 2-го порядка.

Используя обратную функцию шкалы времени и результат интерполяции между отсчетами, была получена новая шкала времени. В результате использования новой шкалы, величины ширины пиков фореграммы изменились. Ширина пиков в старой шкале в диапазоне отсчетов от 1200

до 1450 составляла 16 единиц, а в диапазоне от 16 120 до 16 280 — 23 единицы. После проведения преобразования фореграмм с использованием обратной функции ширина пиков в указанных, а также в других диапазонах оказалась одинаковой и равной 10 единицам.

На рис. 6 показаны фрагменты фореграммы, построенные в исходной (старой) и скорректированной (новой) временных шкалах.

Следует отметить общие особенности и специфику реализации предложенного метода. Основной задачей обработки информации при сиквенс-анализе является обнаружение спектральных пиков — всех пиков в пределах разрешающей способности прибора (секвенатора) и методики эксперимента. При этом основной операцией является преобразование временной шкалы спектров, для осуществления которой необходимо знать параметры пиков.

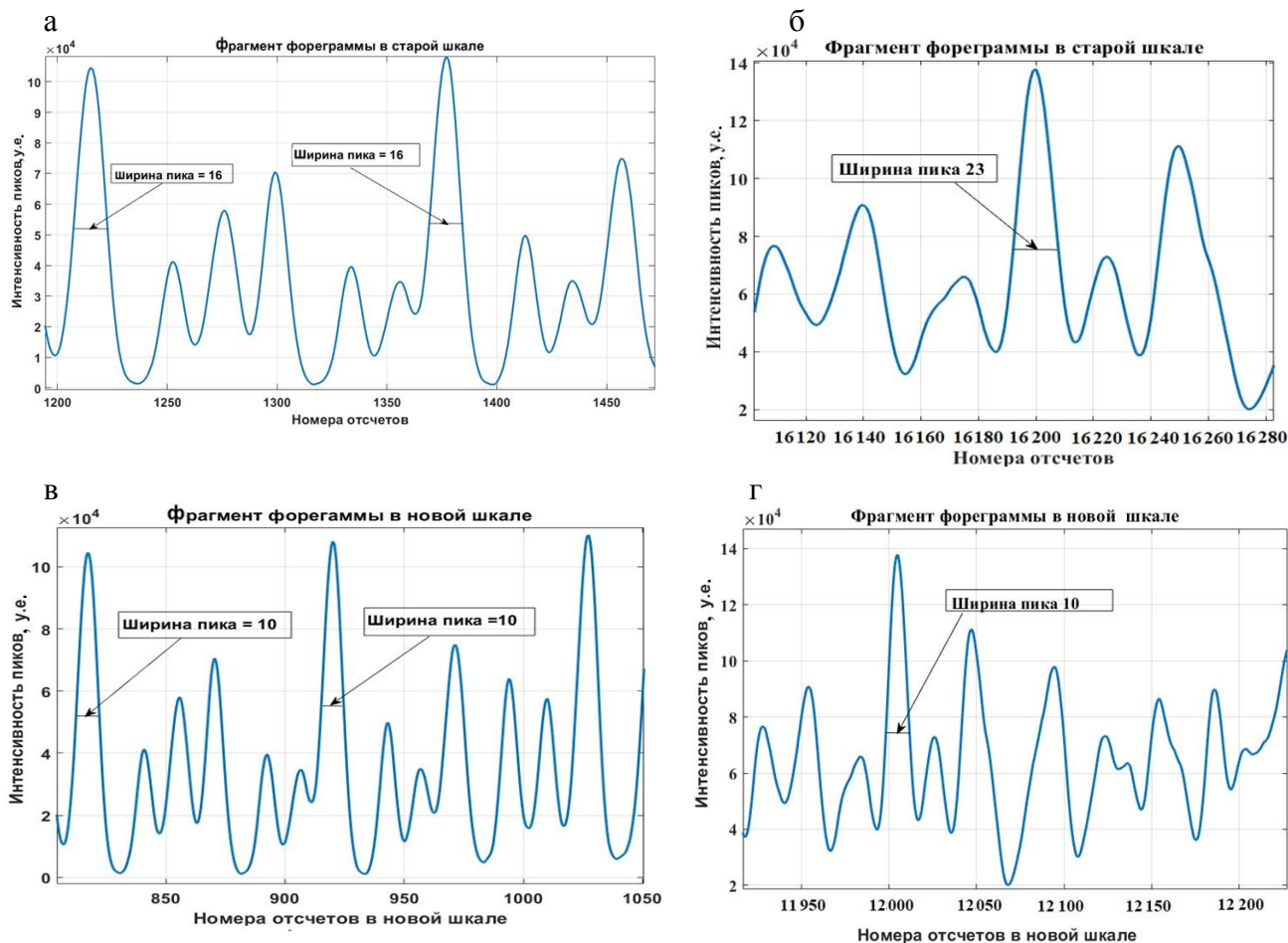


Рис. 6. Фрагменты фореграммы.

а — в начале старой шкалы, б — в конце старой шкалы, в — в начале новой шкалы, г — в конце новой шкалы

Однако для построения шкалирующих функций нет необходимости определять параметры всех пиков спектров. Можно предварительно произвести поиск пиков более грубыми методами (фильтрами), менее критичными к величине ширины пиков. Можно также производить вычисление шкалирующих функций на этапе калибровки прибора с использованием калибровочной пробы, о которой априори известно все.

РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Для достоверной оценки параметров обрабатываемых спектральных сигналов в условиях нестационарных случайных процессов в задачах спектрометрии требуется найти и использовать такой метод коррекции спектров, который позволил бы оценивать параметры спектров с такой же точностью, как в условиях стационарных процессов.

2. Использование обратной функции шкалы времени позволяет скорректировать нелинейную шкалу времени, которая является одним из элементов нестационарности нестационарных случайных процессов в задачах спектрометрии.

3. Для построения обратной функции шкалы времени необходимо произвести обнаружение спектральных пиков, проинтегрировать зависимость расстояний между пиками от времени, а затем в полученной функции заменить зависимость переменной по вертикальной оси от переменной по горизонтальной оси зависимостью переменной по горизонтальной оси от переменной по вертикальной оси.

4. После построения обратной функции отсчета спектра в новой шкале необходимо вычислять при помощи интерполяции между соответствующими соседними отсчетами шкалы времени.

5. Интерполяция между соответствующими соседними отсчетами шкалы времени может осуществляться методом наименьших квадратов при помощи полинома, степень которого выбирается путем минимизации суммы квадратов отклонений между экспериментальными отсчетами времени и аппроксимирующего полинома.

Финансирование

Исследование выполнено в рамках Государственного задания Министерства науки и высшего образования Российской Федерации № 075-00444-25-00 (от 26.12.2024).

СПИСОК ИСТОЧНИКОВ

1. Sanger F., Nicklen S., Coulson A.R. DNA sequencing with chain-terminating inhibitors // Proc. Natl. Acad. Sci. USA. 1977. Vol. 74, no. 12. P. 5463–5467. DOI: 10.1073/pnas.74.12.5463
2. Бочарова Д.В., Алексеев Я.И., Волков А.А., Лавров Г.С., Плугов А.Г., Волков И.А., Чемигов А.А., Бардин Б.В., Курочкин В.Е. Определение максимальной длины ДНК, расшифровываемой с точностью 99 %, в полимере на основе линейного N,N-полидиметилакриламида методом капиллярного гелелектрофореза с лазер-индуцируемой флуоресценцией // Журнал аналитической химии. 2021. Т. 76, № 12. С. 1100–1106. DOI: 10.31857/S0044450221100042
3. Сирвидас С.И., Заруцкий И.В., Ларионов А.М., Манойлов В.В. Обнаружение, разделение и оценка параметров масс-спектрометрических пиков методом свертки экспериментальных данных с производными гауссовых функций // Научное приборостроение. 1999. Т. 9, № 2. С. 71–75. URL: <https://iairas.ru/mag/1999/full2/Art10.pdf>
4. Бардин Б.В. Способ деконволюции спектрометрической информации и обнаружения спектральных пиков // Научное приборостроение. 2017. Т. 27, № 2. С. 75–82. URL: <https://iairas.ru/mag/2017/full2/Art10.pdf>
5. Косарев Е.Л. О пределе сверхразрешения при восстановлении сигналов // Радиотехника и электроника. 1990. Т. 35, № 1. С. 68–87.
6. Jeffreys H., Jeffreys B.S. Inverse Functions in Methods of Mathematical Physics, 2nd ed. Cambridge, Cambridge University Press, 1950. 712 p. URL: <https://djvu.online/file/71tVaT8DGT1p2>
7. Amazigo J.C., Rubinfeld L.A. Advanced Calculus and its Applications to the Engineering and Physical Sciences. New York, Wiley, 1980. 407 p. DOI: 10.1002/zamm.19810611113

Институт аналитического приборостроения РАН,
Санкт-Петербург

Контакты: Бардин Борис Васильевич,
bardin.bv@mail.ru

Материал поступил в редакцию 13.11.2025

A METHOD FOR PROCESSING SIGNALS FROM NON-STATIONARY SPECTROMETRIC PROCESSES IN DNA SEQUENCING USING INVERSE FUNCTIONS

B. V. Bardin, A. Ya. Loginov, V. V. Manoilov, A. I. Petrov

Institute for Analytical Instrumentation of RAS, Saint Petersburg, Russia

The need to analyze non-stationary processes arises in many spectrometry applications. Non-stationary processes are those in which some signal parameters change over time. For a wide variety of devices, time can be understood conditionally. For example, in spectrophotometry, it is the photon wavelength; in mass spectrometry, the molecular mass; in photoelectron spectrometry, the electron binding energy in an atom, etc. As a measure of time in most devices, one can choose a physical characteristic, according to which the spectrum scale — the abscissa axis of the graph, is calibrated. This paper discusses algorithms for processing signals from a non-stationary process, which is a set of signals from Sanger DNA sequencing analysis using capillary electrophoresis. Ideally, in such devices, the width of the spectral peaks and the time interval between the peaks of adjacent nucleotides (in the DNA strand) are the same for all channels. However, in real devices, when analyzing long DNA fragments of approximately 1000 nucleotides, the values of these parameters vary over time. For example, at the end of the analysis cycle, the peak widths increase, while the distances between them decrease, and the spectral peaks practically merge with each other. This paper examines algorithms for converting a non-stationary spectrum into a stationary one by transferring points of the original non-stationary spectrum to the scale of the output stationary spectrum using the inverse function of the time scale. The inverse function allows for correction of the non-linear time scale. The algorithms were tested on mathematical models and real signals. The test results are presented in the paper.

Keywords: nucleic acid sequencing, mathematical processing of non-stationary processes in spectrometry, a use of inverse functions

REFERENCES

1. Sanger F., Nicklen S., Coulson A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, 1977, vol. 74, no. 12, pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463
2. Bocharova D.V., Alekseev Ya.I., Volkov A.A., Lavrov G.S., Plugov A.G., Volkov I.A., Chemigov A.A., Bardin B.V., Kurochkin V.E. [Determination of the maximum length of DNA decoded with 99% accuracy in a polymer based on linear N,N-polydimethylacrylamide by capillary gel electrophoresis with laser-induced fluorescence]. *Zhurnal Analiticheskoy Khimii* [Journal of Analytical Chemistry], 2021, vol. 76, no. 12, pp. 1100–1106. DOI: 10.31857/S0044450221100042 (In Russ.).
3. Sirvydas S.I., Zarutsky I.V., Larionov A.M., Manoilov V.V. [The Convolution of a Signal as Derivatives of Gaussians as an Approach Suitable for Detection, Separation and Estimation of mass spectrometric peaks]. *Nauchnoe Priborostroenie* [Scientific Instrumentation], 1999, vol. 9, no. 2, pp. 71–75. (In Russ.). URL: <https://iairas.ru/mag/1999/full2/Art10.pdf>
4. Bardin B.V. [Way of deconvolution of spectrometer information and detection of spectral peaks]. *Nauchnoe Priborostroenie* [Scientific Instrumentation], 2017, vol. 27, no. 2, pp. 75–82. (In Russ.). URL: <https://iairas.ru/mag/2017/full2/Art10.pdf>
5. Kosarev E.L. [About a superpermission limit at restoration of signals]. *Radiotekhnika i elektronika* [Soviet journal of communications technology and electronics], 1990, no. 1, pp. 68–87. (In Russ.).
6. Jeffreys H. Jeffreys B.S. *Inverse Functions in Methods of Mathematical Physics*, 2nd ed. Cambridge, England: Cambridge University Press, 1950. 712 p. URL: <https://djvu.online/file/71tVaT8DGT1p2>
7. Amazigo J.C., Rubinfeld L.A. *Advanced Calculus and its Applications to the Engineering and Physical Sciences*. New York, Wiley, 1980. 407 p. DOI: 10.1002/zamm.1981061113

INTRODUCTION

The problem of processing signals from non-stationary processes arises in many scientific and technical studies. In spectrometry, the main, "useful" portion of the signal typically represents a time sequence of non-negative spectral peaks. The remaining signal components, particularly the background component, noise, and interference, are usually removed at the initial stages of processing. However, at subsequent stages, the need often arises to analyze non-stationary processes. These are processes in which some signal parameters change over time. Time here can be understood conditionally. Generally, it is the physical characteristic by which the spectral scale — the abscissa axis of the graph — is calibrated. For example, in photospectrometry, this is the photon wavelength; in mass spectrometry, this is the molecular mass; in photoelectron spectrometry, this is the electron binding energy in an atom, and so on.

The most typical example of such spectral signals is the set of signals from Sanger sequencing analysis of DNA fragments using capillary electrophoresis [1, 2]. A phoregram — the data recorded as a result of electrophoresis — consists of four spectra, or four color channels, on a single time scale. The main signal parameters — the width of the spectral peaks and the time interval between the peaks of adjacent nucleotides (in the DNA strand) — are the same for all channels. However, when analyzing long DNA fragments, approximately 1000 nucleotides, the parameter values change significantly over time. Thus, at the end of the analysis cycle, the peak width increases and the distance between them decreases, and the spectral peaks practically merge into each other.

The parameters of most spectral processing operations (filtering, etc.) must be matched to the parameters of the useful signals and noise. For example, the expression for describing a peaking filter [3–5] used to separate merging peaks includes an image of the desired useful peak, the parameters of which are the peak width and the parameters or mathematical expression of the peak shape (e.g., a Gaussian). The image parameters directly determine the accuracy and correctness of separating the merging signal peaks.

Correctness can be illustrated as follows. If we analyze a burst of two identical fused peaks and if the signal peak width and the filter image are equal, then two peaks are typically identified (if we are working within the resolution of the method). If, however, a filter image with a width half the signal peak width at a given spectral point is used (which is quite possible), then three peaks are typically obtained.

To reliably estimate the parameters of the processed spectral signals of non-stationary random processes, it is necessary to find and use a spectral correction method that would allow us to estimate the

spectral parameters with the same accuracy as under stationary processes. This paper proposes one of the possible methods for estimating the parameters of the spectral signals of non-stationary processes, which is based on the use of inverse functions [6], [7, pp. 103–120]. The method has been tested in signal processing tasks in the software of a domestic DNA sequencer.

PROPOSED SOLUTION

Transformation of a non-stationary spectrum into a stationary one

Under certain conditions, such a transformation can be accomplished by transferring points (samples) of the original non-stationary spectrum to the scale of the output stationary spectrum with a time shift proportional to the difference of the non-stationarity parameter interval from a certain value specified as the corresponding constant interval of the output spectrum. This latter interval, in a particular case, can be equal to the average value of the interval in the original spectrum (then the duration of the resulting spectrum will be equal to the duration of the original) or chosen for some other reason.

Time scale transformation

The problem is solved under the following conditions:

- *The non-stationarity parameter*, which must be made constant as a result of the transformation, is a positive, non-zero time interval through the entire scale. For example, a parameter such as the width of a spectral peak cannot be negative or zero — this has no physical meaning.

- *The scaling function* describing the change in the parameter along the scale is slow and smooth (see below). For example, for a spectrum of tens of thousands of samples, such a function can be quite acceptably represented by a fifth-order polynomial, given that the width of the spectral peaks (the non-stationarity parameter) is typically on the order of 5–30 samples.

The authors provide numerical parameter estimates for the problem of capillary electrophoresis for DNA analysis.

The following notations are used in this paper:

t is the "physical" time, which determines the scale of the original, recorded spectrum*;

*) This notation is used here to avoid confusion with, for example, such established terms as "real" time (online). It should be noted that spectrometric information is typically processed in "non-real" time (offline), when the entire spectrum has already been fully accumulated and recorded.

τ is the "virtual" time, which determines the scale of the "virtual," ideal spectrum, where the interval of the nonstationarity parameter is constant;

δ is the nonstationarity parameter on the t scale, for example, the width of a spectral peak — a time interval;

Δ is the value of the nonstationarity parameter on the "ideal," virtual scale τ — a constant time interval.

Let's imagine the following mental model of the system in which the problem under discussion must be solved.

The input process is an "ideal" spectrum with a constant width of spectral peaks Δ over the entire time τ . However, when it is recorded by the instrument, a distortion of the time scale t of the output recorded spectrum occurs. For example, due to the "floating" frequency of the master oscillator, which generates the samples of digitalization of the output spectrum. Accordingly, the width of the spectral peaks $\delta(t)$ on this scale changes in the number of samples.

To justify the choice of this example, we say that it is the most illustrative, although also the least probable, since master oscillators are perhaps the most stable and precise electronic components.

The scale distortion is described by a scaling function $t = f(\tau)$ (see Fig. 1).

The problem is to reconstruct the original ideal spectrum from the recorded spectrum on the t scale.

If the function $f(\tau)$ is strictly monotonic, then there exists an inverse function $\tau = f^{-1}(t)$, which is also strictly monotonic. And from the theory of inverse functions [6, 7], it is known that

$$f^{-1}(f(\tau)) = \tau, \quad f(f^{-1}(t)) = t. \quad (1)$$

Thus, in order to reconstruct, and in reality create, the "ideal" τ scale in our model, it is necessary to subject the physical t scale to an inverse transformation by the function $f^{-1}(t)$.

In reality, we have a recorded physical spectrum on the t scale. This spectrum can be processed, spectral peaks can be found, and the peak width versus time dependence $\delta(t)$ can be plotted from them (Fig. 1).

Fig. 1. Explanations for plotting a scale using an inverse function.

a — scaling and inverse functions, δ — peak width versus time

In Fig. 1, the quantities Δ and δ are shown illustratively to make things clear; in reality, these quantities are 2–3 orders of magnitude shorter than the scale length. Therefore, they are quite acceptable to use as differentials. Then, from Fig. 1 derivative of τ with respect to t will be

$$\frac{d\tau(t)}{dt} = \frac{\Delta}{\delta(t)}. \quad (2)$$

The function $f^{-1}(t)$ is calculated by integrating its derivative (3)

$$f^{-1}(t) = \tau(t) = \Delta \int_0^t \frac{1}{\delta(t)} dt. \quad (3)$$

The function $1/\delta(t)$ is a nonzero positive function, and its integral will be strictly monotone. Therefore, the above condition for the function $f^{-1}(t)$ to be strictly monotone is satisfied.

The function $f(\tau)$ can be calculated in a similar manner. However, the simplest and most reliable method (in terms of eliminating the possible accumulation of calculation errors during integration) is to form f as a mirror image (or the result of a 180-degree rotation) of the function f^{-1} about the $\tau = t$ axis. The point is that the requirement for the accuracy of calculating these functions is, generally speaking, quite moderate, but the accuracy of their relationship and the accuracy of the transformations (1) must be absolute.

After the τ scale is created, the spectral data from the t scale are transferred to it. To do this, τ_i samples are defined on the τ scale, typically with a constant step. For each of these samples, the corresponding time t_j is calculated using the function $f^{-1}(t)$, and the spectral data value of $D[t(t_j)]$ corresponding to this time is transferred to the sample of the i -th τ scale — $D[\tau(\tau_i)]$. It should be noted that the time t_j will almost always not exactly coincide with the time of any sample on the t scale. Therefore, the value of $D[t(t_j)]$ must be calculated by interpolation between corresponding adjacent samples on the t scale.

Example of time scale transformation for a phoregram spectrum model containing peaks of varying widths

The time scale transformation method discussed in the previous subsection was first tested using a phoregram model of 25 000 samples, containing 160 Gaussian peaks. The peaks were equally spaced, but each peak had a different width. The peak width changed linearly over time.

It should be noted that in discrete mathematics and programming, elements of data arrays are identified by their numbers within the array. Therefore, for simplicity, in the figures below, the t and τ scales are digitized in samples and have the same notations in the figures. The peak width scales are also digitized in terms of the number of samples. These digitizations will correspond to seconds if the distance between samples is 1 s. However, the experimental data used in this study were obtained from a sequencer with

a sample interval set to 0.25 s. This means that the value of these digitizations in seconds will be four times smaller.

Figure 2, a, shows a fragment of the spectrum model, indicating the peak widths for the beginning and end of the scale before applying the scale linearization algorithm. Figure 2, б, shows the peak widths as a function of time for the spectrum model presented in Figure 2, a. Figure 2, в, shows a fragment of the spectrum model, demonstrating the peak widths for the beginning and end of the scale after reconstructing the time scale using the inverse function discussed above. Figure 2, г, shows the peak widths for the original nonlinear scale and for the scale reconstructed using the inverse function. Interpolation based on second-order polynomials was used to reconstruct the time scale.

Fig. 2. Fragments of the phoregram model and peak width dependences.

a — fragment of the phoregram model with peaks of various widths at the beginning and end of the scale, the peak width is given as the number of samples; б — dependence of the peak width on the sample numbers;

в — fragment of the spectrum model indicating the peak widths for the beginning and end of the scale after restoring its linearity;

г — dependences of the peak width w for the original nonlinear scale and for the scale restored using the inverse function

As can be seen from Fig. 2, г, the dependence of the peak width on time in the range from 0 to 20 000 is a line close to a straight line, parallel to the horizontal axis.

The average peak width in this region is 23.9, and the standard deviation is 0.1. A slight increase in peak width due to edge effects can be seen in the region from 20 000 to 25 000. The average peak width across the entire spectrum model range is 24.0, and the standard deviation is 0.4.

A study of the method for restoring scale linearity using the inverse function, conducted using the model, demonstrated its effectiveness for processing real phoregrams.

Example of time scale linearization for a phoregram obtained using a DNA sequencer

Figures 3, a and б show fragments of a phoregrams obtained using a Nanofor-05 sequencer: a — fragment at the beginning of the scale, б — fragment at the end of the scale. The peak width has increased at the end of the scale.

Fig. 3. Fragments of the phoregram: a — at the beginning of the scale, б — at the end of the scale.

Figure 4 shows the dependence of peak width on the number of samples at half-height of the peak. The solid line shows the approximation of this dependence by a fifth-degree polynomial. This approximation was performed using the least-squares method to construct the regression line. The parameters of the approximation function characterize the nonlinearity of the time scale.

Fig. 4. Dependence of peak width on sample numbers.

Solid line - polynomial fit

Figure 5 shows the direct and inverse functions of the phoregram time scale. The lower part of the figure shows the function that results from integrating the peak width approximation function.

Fig. 5. Direct $t = f(\tau)$ and inverse $\tau = f^{-1}(t)$ functions of the phoregram time scale

The integral of the interpeak time approximation function characterizes the time scale and is $t = f(\tau)$. Following the algorithm discussed in the previous section, we find the inverse function $\tau = f^{-1}(t)$, replacing the dependence of the vertical axis variable on the horizontal axis variable with the dependence of the horizontal axis variable on the vertical axis variable.

The value of $D[t(t_i)]$ was calculated as described above, using a second-order polynomial to interpolate $D[t(t)]$ between adjacent samples.

Using the inverse time scale function and the interpolation result between samples, a new time scale was obtained. As a result of using the new peak width scale, the phoregrams changed. The peak width on the old time scale was 16 units in the 1200 to 1450 sample range and 23 units in the 16 120 to 16 280 range. After transforming the phoregrams using the inverse function, the peak widths in these and other ranges were the same and equal to 10 units.

Figure 6 shows fragments of a phoregram constructed in the original (old) and corrected (new) time scales.

Fig. 6. Fragments of the phoregram.

a — at the beginning of the old scale, $\bar{o}$ — at the end of the old scale, b — at the beginning of the new scale, r — at the end of the new scale

The general features and implementation specifics of the proposed method should be noted. The primary objective of data processing in sequence analysis is to detect spectral peaks — all peaks within the resolution of the instrument (sequencer) and the experimental methodology. The key operation is the transformation of the spectral time scale, which requires knowledge of the peak parameters.

However, to construct scaling functions, it is not necessary to determine the parameters of all spectral peaks. A preliminary peak search can be performed using more coarse methods (filters) that are less sensitive to peak width. Scaling functions can also be calculated during the instrument calibration stage using a calibration sample, about which everything is known a priori.

RESULTS AND CONCLUSIONS

1. To reliably estimate the parameters of processed spectral signals under non-stationary random

processes in spectrometry problems, it is necessary to find and use a spectral correction method that would allow estimating spectral parameters with the same accuracy as under stationary processes.

2. Using the inverse time scale function allows for correction of the nonlinear time scale, which is one of the elements of non-stationarity of non-stationary random processes in spectrometry problems.

3. To construct the inverse function of the time scale, it is necessary to detect spectral peaks, integrate the dependence of the distances between peaks on time, and then replace the dependence of the variable along the vertical axis on the variable along the horizontal axis in the resulting function with the dependence of the variable along the horizontal axis on the variable along the vertical axis.

4. After constructing the inverse function, the spectrum points in the new scale must be calculated using interpolation between the corresponding adjacent points of the time scale.

5. Interpolation between the corresponding adjacent readings of the time scale can be performed by the least-squares method using a polynomial which degree is selected by minimizing the sum of the squares of the deviations between the experimental time points and the approximating polynomial.