

УДК 543.426; 543.9

© Ю. В. Белов, И. А. Леонтьев, А. И. Петров, В. Е. Курочкин

КОРРЕКЦИЯ БАЗОВОЙ ЛИНИИ СИГНАЛОВ ФЛУОРЕСЦЕНТНОГО ДЕТЕКТОРА ГЕНЕТИЧЕСКОГО АНАЛИЗАТОРА

Выполнено имитационное моделирование сигналов фрагментов ДНК флуоресцентного детектора генетического анализатора и определены погрешности вычисления базовой линии. Предложен способ вычисления базовой линии, основанный на выделении минимальных значений сигнала на выбранном участке, обладающий значительно меньшими погрешностями по сравнению с известным способом медианной фильтрации.

Кл. сл.: ДНК, генетический анализатор, флуоресцентная детекция

ВВЕДЕНИЕ

Целью генетического анализа является разделение (секвенирование) ДНК на фрагменты и определение длины фрагментов и последовательности нуклеотидов в этих фрагментах [1, 2]. В генетических анализаторах разделение фрагментов ДНК происходит в капилляре под действием электрического поля. При достижении флуоресцентно-мечеными фрагментами ДНК оптического окна, находящегося вблизи конца капилляра (более короткие фрагменты достигают его быстрее, а более длинные — медленнее), лазером производится возбуждение флуоресцентных красителей в составе фрагментов ДНК, а регистрация флуоресценции — детектором. Сигналы флуоресценции получаются за счет возбуждения 5 флуоресцентных красителей, регистрируются 5-канальным детектором в виде 5 цифровых последовательностей и индицируются в виде 5 графиков. При этом каждому цифровому отсчету соответствует точка графика, имеющая по горизонтальной оси (оси времени) номер отсчета, а по вертикальной оси — значение интенсивности флуоресценции в относительных единицах (о. е.). Четыре канала используются для секвенирования ДНК. Пятый канал детектора используется для калибровки длин фрагментов.

В 2011 г. Институтом аналитического приборостроения РАН (Санкт-Петербург) совместно с ЗАО "Синтол" (Москва) был разработан и успешно испытан макет генетического анализатора [3].

Обычно обработка сигналов детектора начинается с устранения начального смещения и дрейфа уровня базовой линии отдельно в каждом цветовом канале. В генетическом анализаторе [3] был

применен традиционный способ вычисления базовой линии — медианная фильтрация. Медианная фильтрация хорошо устраняет одиночные пики, однако при анализе последовательности ДНК большое количество пиков находится на близком расстоянии друг к другу, поэтому, если применить сильное медианное сглаживание таких исходных данных, то сглаженная линия в процессе вычисления будет накапливать ошибки за счет частичного интегрирования пиков. После вычисления базовая линия вычитается из исходного ("сырого") сигнала, при этом за счет ошибок вычисления базовой линии сигналы ДНК искажаются.

В настоящей статье предложен способ вычисления базовой линии, обладающий значительно меньшими погрешностями. Этот способ использован при разработке 8-капиллярного генетического анализатора, выполняемой в рамках Государственного контракта с Министерством образования и науки РФ [4]. Этот генетический анализатор по своим параметрам и пользовательским характеристикам не уступает лучшим импортным аналогам и обладает многими конкурентными преимуществами.

МОДЕЛИРОВАНИЕ СИГНАЛОВ ФЛУОРЕСЦЕНТНОГО ДЕТЕКТОРА

С целью определения погрешностей вычисления базовой линии выполнено имитационное моделирование сигналов фрагментов ДНК в одном из каналов флуоресцентного детектора в виде случайной последовательности пиков, в которой некоторые пики произвольно располагаются поодиночке, а другие — группами, включающими 2 и 3 соседних пика. Моделирование имеет преимущество по сравнению с реальными сигналами, по-

сколько известны параметры отдельных составляющих сигналов: пиков, базовой линии и шума (анализ влияния шума будет выполнен в следующей статье). При реализации численных экспериментов выбраны следующие параметры сигналов фрагментов ДНК:

- форма пиков определяется функцией Гаусса;
- высота пиков $M = 1000$ о. е.;
- $e = 2.72$ — приближенное значение;
- график имеет 5 участков по 300 точек, ширина пиков на уровне 0.6 на участках соответственно равна $2\sigma = 4, 6, 8, 9$ и 10 точек;
- минимальное расстояние между пиками 10 точек;

– положение центра пика: 41, 81, 91, 141, 151, 161, 241 (точка).

При общей длине последовательности 1501 точка и минимальном расстоянии между пиками, равном 10 точкам, суммарное количество пиков 4-канального детектора равно 150. Общее количество пиков в первом канале, равное 35, выбрано при условии, что остальные 115 пиков можно почти равномерно распределить при моделировании сигналов во 2, 3 и 4-м каналах детектора.

Переменная ширина пиков соответствует реальному ухудшению разрешения в конце электрофореза.

Табл. 1. Результаты вычисления модельных сигналов в эл-й таблице (первые 4 строки из 1501 строки)

№ строки	A	B	C	D	E	F	G
1	0	1,22E-84	0,00E+00	1,22E-84	5,99E-10	4,07E+01	25,00
2	1	2,39E-80	1,67E-01	1,67E-01	5,99E-10	4,09E+01	25,17
3	2	3,63E-76	3,34E-01	3,34E-01	5,99E-10	4,12E+01	25,33
4	3	4,31E-72	5,01E-01	5,01E-01	5,99E-10	4,14E+01	25,50

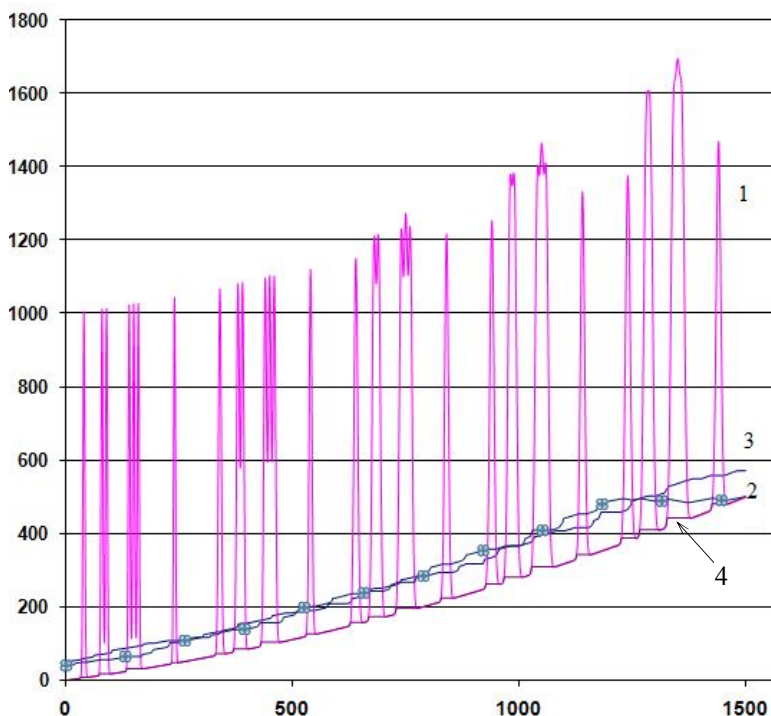


Рис. 1. Модельный сигнал и вычисленные базовые линии.

1 — модельная последовательность пиков; 2 — результат медианного сглаживания последовательности пиков с дрейфом базовой линии участка длиной 1501 точка; 3 — результат сглаживания медианой последовательности пиков при удлинении базовой линии; 4 — базовая линия, вычисленная путем выделения минимальных значений сигнала на выбранном участке (имеет ступенчатый вид). Горизонтальная ось — номер отсчета, вертикальная ось — значение интенсивности флуоресценции в относительных единицах (о. е.)

В терминах программы Excel функция выбранной последовательности пиков в столбце В табл. 1 имеет вид:

$$B1 = M \cdot 2,72^{-(X1-A1)^2 / (2 \cdot \sigma^2)} + M \cdot 2,72^{-(X2-A1)^2 / (2 \cdot \sigma^2)} + \dots,$$

где $X1 = 41$ — положение центра первого пика;
 $X2 = 81$ — положение центра второго пика и т. д.;
 $A1 = 0$ — порядковый номер ячейки в столбце А;

$B1$ — относительная интенсивность сигнала флуоресценции в первой ячейке столбца В.

Значения ячеек $B2, B3, \dots, B1501$ находятся соответственно при $A2 = 1, A3 = 2, \dots, A1501 = 1500$.

Дрейф базовой линии моделировался в столбце С полиномом второй степени:

$$C1 = m \cdot A1 + n \cdot A1^2,$$

где m и n — коэффициенты, выбранные из условий $m \cdot A1501 = 250$ и $n \cdot A1501^2 = 250$ (суммарный дрейф в конце графика равен 500 о. е.).

Модельная последовательность пиков, объединенная с модельной базовой линией, приведена в столбце D табл. 1 и на рис. 1 — кривая 1:

$$D1 = B1 + C1.$$

ВЫЧИСЛЕНИЕ БАЗОВОЙ ЛИНИИ МОДЕЛЬНЫХ СИГНАЛОВ С ИСПОЛЬЗОВАНИЕМ СПОСОБА МЕДИАННОЙ ФИЛЬТРАЦИИ

Вычисление базовой линии способом медианной фильтрации последовательности пиков с базовой линией без дрейфа выполнена в столбце Е по формуле:

$$E1 = \text{МЕДИАНА}(B1:B301).$$

Медианное сглаживание имеет существенную максимальную погрешность вычисления базовой линии (более 20 о. е.) даже при идеальной исходной базовой линии и оптимальной длине участка сглаживания (1:301).

Результат медианного сглаживания последовательности пиков с дрейфом базовой линии приведен в столбце F и на рис. 1 — кривая 2:

$$F1 = \text{МЕДИАНА}(D1:D301).$$

Для устранения краевого эффекта предлагается удлинить базовую линию и продлить ее график влево и вправо на 150 точек.

Результат сглаживания медианой последовательности пиков при удлинении базовой линии

приведен в столбце G. Этот прием можно применить с учетом того, что анализ ДНК длится приблизительно 100 мин, при этом полезный сигнал (с пиками) начинается примерно после 15-й мин, а регистрация базовой линии продолжается после выхода всех пиков еще некоторое время. Однако такой прием не дает существенного уменьшения погрешностей вычисления базовой линии (рис. 1, кривая 3). Максимальная погрешность (отличие от исходной базовой линии) составляет 112 о. е., или 22.4 % от значения максимального дрейфа.

СПОСОБ УМЕНЬШЕНИЯ ПОГРЕШНОСТЕЙ ВЫЧИСЛЕНИЯ БАЗОВОЙ ЛИНИИ

С целью значительного уменьшения погрешностей предложен способ вычисления базовой линии, основанный на выделении минимальных значений сигнала на выбранном участке.

Очевидно, что участок сглаживания нужно выбрать так, чтобы охватывать пики максимальной ширины. Например, если положения центров пиков соответствуют 141-, 151- и 161-й точкам, то можно использовать участок сглаживания длиной 60 точек. При этом базовая линия последовательности пиков без дрейфа восстанавливается практически идеально.

Для построения базовой линии предложенным способом при дрейфе базовой линии использована следующая функция в столбце H табл. 2:

$$H1 = \text{МИН}(D1:D60).$$

На графике рис. 1, кривая 4, видно, что погрешность вычисления предложенным способом значительно меньше, чем при использовании медианы (рис. 1, кривая 3). Максимальная погрешность до 25 о. е. наблюдается (см. рис. 1 (4)) на интервалах, соответствующих широким пикам.

Дальнейшее уменьшение максимальной погрешности вычисленной базовой линии достигается аппроксимацией этой линии монотонной функцией. Аппроксимирующая функция вычисленной базовой линии может быть выражена в виде полинома третьей степени в столбце J в следующем виде:

$$J1 = \$K\$1 + \$K\$2 \cdot A1 + \$K\$3 \cdot (A1)^2 + \$K\$4 \cdot (A1)^3,$$

где исходные величины выбираются из условий: $\$K\$1, \$K\3 и $\$K\4 — малые величины (равны нулю), $\$K\$2 = H2 - H1 = 0,17$ — ориентировочное единичное приращение (наклон графика на рис. 2).

В столбце L вычисляются квадраты отклонений

$$L1 = (J1 - H1)^2,$$

Табл. 2. Результаты вычисления в эл-й таблице аппроксимирующей функции базовой линии (первые 4 строки)

Н	К	Ј	Л	М
0,00	3,37632	3,38	11,3995	35831,25
0,17	0,168004	3,54	11,4093	—
0,33	0,00011	3,71	11,4191	—
0,50	5,21E-10	3,88	11,4288	—

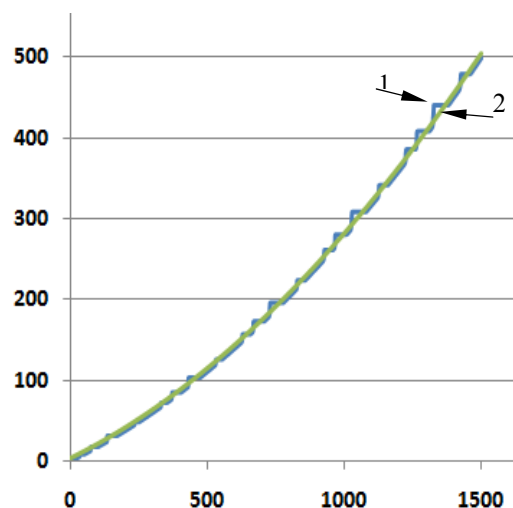


Рис. 2. График базовой линии.

1 — базовая линия, вычисленная путем выделения минимальных значений сигнала на выбранных участках (имеет вид ступенек); 2 — аппроксимирующая функция (монотонная). Обозначения осей аналогичны рис. 1

в столбце М — сумма квадратов отклонений

$$M1 = \text{СУММ}(L1:L1501).$$

Величины в столбце К подбираются с целью максимального совпадения графиков на рис. 2, а затем уточняются по методу наименьших квадратов. Для этого используется меню "Данные\Анализ\Поиск решения" и выполняются многократно следующие действия: "Установить целевую ячейку M1", "Равной минимальному значению", "Изменяя ячейки \$K\$1", "Изменяя ячейки \$K\$2" и т. д. Пример вычисления аппроксимирующей функции базовой линии приведен в табл. 2.

График вычисленной базовой линии, аналогичный рис. 1 (4), и график аппроксимирующей его функции приведены на рис. 2 (1, 2). Максимальная погрешность 5.36 о. е. наблюдается в конце графика.

В случае отрицательного дрейфа, равного –500 о. е., значения максимальных погрешностей вычисления при дрейфе базовой линии составляют соответственно 29.1 (без аппроксимации) и 22.0 о. е. (с аппроксимацией). Последнее значение соответствует относительной ошибке около 4.4 % от максимального дрейфа нулевой линии, что соответствует уменьшению дрейфа более чем в 20 раз.

Дальнейшее значительное уменьшение погрешности вычисления базовой линии достигается повторным использованием способа выделения минимальных значений сигнала и вычисления ап-

проксимирующей функции. Для последнего численного примера в случае отрицательного дрейфа максимальная ошибка вычисления базовой линии находится в пределах до –0.42 о. е., или 0.1 % от максимального дрейфа базовой линии.

ЗАКЛЮЧЕНИЕ

С целью определения погрешностей вычисления базовой линии выполнено имитационное моделирование сигналов фрагментов ДНК в каналах флуоресцентного детектора в виде последовательностей одиночных и сгруппированных пиков, форма пиков определяется функцией Гаусса.

Предложен способ вычисления базовой линии, основанный на выделении минимальных значений сигнала на выбранном участке и аппроксимации вычисленной линии монотонной функцией. Предложенный способ обладает значительно меньшими погрешностями (в рассмотренном примере в 200 раз) по сравнению с традиционным способом вычисления — медианной фильтрацией. Дополнительное преимущество предложенного способа заключается в меньшем искажении динамики базовой линии за счет уменьшения участка сглаживания до 60 точек по сравнению с 300 точками при медианной фильтрации.

Работа выполнена при поддержке Министерства образования и науки Российской Федерации в рамках Федеральной целевой программы "Исследования и раз-

работки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы" и опытно-конструкторской работы "Разработка генетического анализатора для секвенирования и фрагментного анализа ДНК" (шифр заявки "2011-2.2-522-014-001", Государственный контракт № 16.522.12.2014 от 10 октября 2011 г.).

конференция "Аналитические приборы", 25–30 июня 2012, Санкт-Петербург. Тезисы докладов. С. 104.

4. Алексеев Я.И., Белов Ю.В., Малюченко О.П. и др. Генетический анализатор для фрагментного анализа ДНК. // Научное приборостроение. 2012. Т. 22, № 4. С. 86–92.

СПИСОК ЛИТЕРАТУРЫ

1. Бельский Б.Г., Козулин Р.А., Курочкин В.Е., Золотарев В.М. Аналитический метод определения последовательности ДНК (секвенирование) по спектральным флуоресцентным меткам // Оптический журнал. 2003. Т. 70, № 1. С. 65–68.
2. Алексеев Я.И., Белов Ю.В., Варламов Д.А. и др. Прибор капиллярного электрофореза "НАНОФОР 03-С" для определения последовательности и фрагментного анализа нуклеиновых кислот // Материалы 4-го съезда Общества биотехнологов России им. Ю.А.Овчинникова, 06–07 декабря 2006, Москва. М.: Макс пресс., 2006. С. 10–11.
3. Курочкин В.Е., Алексеев Я.И., Белов Ю.В. и др. Генетический анализатор для секвенирования и фрагментного анализа ДНК // 4-я Всероссийская

**Институт аналитического приборостроения РАН,
г. Санкт-Петербург**

Контакты: Белов Юрий Васильевич,
bel3838@mail.ru

Материал поступил в редакцию 2.04.2013

SIGNAL BASELINE CORRECTION OF THE FLUORESCENT DETECTOR OF GENETIC ANALYZER

Yu. V. Belov, I. A. Leontyev, A. I. Petrov, V. E. Kurochkin

Institute for Analytical Instrumentation of RAS, Saint-Petersburg

Imitating simulation of DNA fragments signals of genetic analyzer of fluorescent detector was carried out and baseline computation errors were determined. Proposed baseline calculation method, based on signal minimum values detection in a selected area. This method has significantly fewer calculating errors in comparison with the known method of median filtering.

Keywords: DNA, genetic analyser, fluorescence detection