

Сравнение методов определения базовой линии сигналов флуоресцентного детектора генетического анализатора

Ю.В. Батов¹, Д.А. Пузько¹, А.И. Петров², В.В. Давыдов^{1,3}, В.Ю. Рудь³

¹Высшая школа прикладной физики и космических технологий, Санкт-Петербургский Политехнический университет Петра Великого, ул. Политехническая 29, Санкт-Петербург, Россия, 195251

²Институт Аналитического Приборостроения Российской Академии Наук, Рижский пр. 26, Санкт-Петербург, Россия, 190103

³Всероссийский научно-исследовательский институт фитопатологии, ул. Институт 5, р.п. Большие Вяземы, Россия, 143050

Аннотация. В статье рассмотрены методы определения базовой линии сигналов флуоресцентного детектора генетического анализатора. Выполнено моделирование сигнала, схожего по своим характеристикам с реальными сигналами. Проведено сравнение методов по критерию среднеквадратичной ошибки и скорости вычисления на модельных сигналах. Выполнен качественный анализ эффективности методов на реальных сигналах. Представлены теоретические выкладки по данным методам и результаты исследований.

1. Введение

Одним из методов, которые используются в генетическом анализе, является секвенирование ДНК по Сэнгеру [1-3]. Этот метод обладает рядом преимуществ по точности информативности исследований по сравнению с другими методами, например, ядерным магнитным резонансом [4 - 6]. В процессе секвенирования ДНК разделяется на фрагменты с шагом в 1 нуклеотид. На конец каждого фрагмента помещается флуоресцентный краситель в зависимости от последнего нуклеотида в фрагменте (всего используются 4 красителя, один для каждого нуклеотида). В капилляре под действием электрического поля происходит разделение фрагментов ДНК [1-3]. В конце капилляра находится лазер, который возбуждает флуоресцентный краситель, это возбуждение регистрирует специальный детектор в отличие от классических приборов, применяемых в оптических исследованиях [6-9].

Для секвенирования используются 4 канала. Данные с выхода детектора оцифровываются в виде 4 цифровых последовательностей. По горизонтальной оси откладываются номера отсчетов сигнала, по вертикальной – интенсивность света в относительных единицах.

Обработка сигналов детектора является нетривиальной задачей. Ввиду аппаратных свойств, наличия шумов и иных факторов, воздействующих на лазерное детектирование, появляются шумы, базовая линия и слияние пиков. Обычно первой задачей ставится удаление базовой линии. Существует множество методов для определения базовой линии. Одним из самых простых – это медианная фильтрация. Данный метод применяется в генетическом анализаторе, разработанном ИАП РАН совместно с ЗАО «Синтол» [11]. Однако при использовании данного метода возникают значительные погрешности, способные исказить исходные данные.

В работе рассмотрены альтернативные методы по устранению базовой линии, обладающие значительно меньшими погрешностями, в сравнении с медианной фильтрацией. Эти методы применены к модельному сигналу и к реальным данным. Проведено сравнение по критерию минимальной среднеквадратичной ошибки (СКО) предложенных методов.

2. Моделирование идеального сигнала с выхода детектора генетического анализатора

Для вычисления среднеквадратичной ошибки найденной базовой линии необходимо знать заданную базовую линию. Для этого было проведено моделирование сигнала с теми особенностями, которые были выявлены на реальных данных.

Моделирование сигнала производилось по методике, описанной в [12]. Генерируется массив длиной 600, каждый элемент которого является случайным числом от 1 до 4 (каждое соответствует своему нуклеотиду). Задаются массы нуклеотидов и красителей [12]. Создаются фрагменты из нуклеотидов с шагом 1, то есть первый фрагмент состоит из первого нуклеотида, второй из первых двух и т.д. К вычисленной массе каждого фрагмента добавляется масса красителя в зависимости от типа последнего нуклеотида в фрагменте.

$$m_i = \sum_{j=1}^i m_{nj} + m_{kpi} \quad (1)$$

где m_i – масса i -го фрагмента, m_{nj} – масса j -го нуклеотида, m_{kpi} – масса красителя, соответствующая i -му нуклеотиду.

Считая движение фрагментов в капилляре равноускоренным и зная массу, находим время, за которое приходит каждый фрагмент. В эти моменты образуются пики в одном из четырех сигналов. Вид пика приближенно считаем гауссовским. Ширину гауссовского пика увеличивали с увеличением времени, такой эффект наблюдается в реальных сигналах. Также было замечено, что в реальных сигналах амплитуда пика с увеличением времени уменьшается, однако присутствуют и пики, амплитуды которых меняются случайным образом. В модельном сигнале амплитуда A генерируется следующим образом:

$$A = at + x, \quad (2)$$

где a характеризует общую линейную зависимость, t время, x значение из равномерного распределения от 0 до 1.

Таким образом, общее увеличение амплитуды сохраняет линейный характер, но каждое конкретное значение являлось случайной величиной. Окончательное уравнение идеального модельного сигнала s можно представить в следующем виде:

$$s(x) = \sum_i A_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x-m_i)^2}{2\sigma_i^2}\right), \quad (3)$$

где σ_i – ширина гауссовского пика (в терминах нормального распределения (НР) – дисперсия), m_i – масса i -го фрагмента (в терминах НР – мат. ожидание).

К полученному модельному сигналу (3) был добавлен шум и базовая линия. Базовая линия $B(t)$ была задана в виде:

$$B(t) = 0.3 \cdot 10^4 \left(\sin \frac{x}{10} + \sin \frac{x}{20} + \sin \frac{x}{40} \right). \quad (4)$$

На базовую линию был наложен шум путем добавления к каждому значению базовой линии значения из случайной равномерно распределенной величины от 0 до 1. На рисунке 1 изображен вид модельного сигнала. В качестве среды для моделирования сигнала выбран MATLAB.

3. Методы определения базовой линии

В современном генетическом анализе для удаления базовой линии используются различные методы. Часто применяемыми методами являются метод асимметричных наименьших квадратов (АНК) [13] и метод, основанный на фильтре Савицкого-Голея [14]. Данные методы были сравнены с методом медианной фильтрации и был найден оптимальный по критерию СКО и времени вычисления.

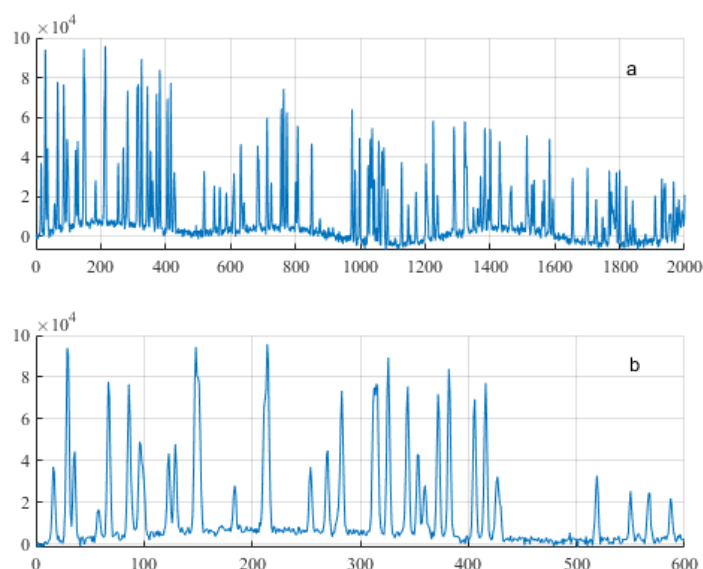


Рисунок 1. Моделирование сигнала с выхода детектора. График а) большой масштаб. График б) фрагмент сигнала а).

3.1 Метод асимметричных наименьших квадратов

Метод асимметричных наименьших квадратов основывается на методе Уиттекера с весовыми коэффициентами [15]. Сглаживание Уиттекера определяется следующей формулой:

$$S = \sum_i (y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2, \quad (5)$$

$$\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2}, \quad (6)$$

где y_i - отсчеты реальной базовой линии, z_i - отсчеты аппроксимированной базовой линии, λ - весовой коэффициент. Первое слагаемое в (5) является среднеквадратичным отклонением аппроксимированной базовой линии от реальной, второе слагаемое определяет сглаживание. Найдем минимум значения S . Перепишем выражение в матричном виде:

$$S = (y - z)'(y - z) + \lambda z' D' D z, \quad (7)$$

$$y = (y_1, \dots, y_n)', z = (z_1, \dots, z_n)', \quad (8)$$

$$D = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots \\ 0 & 1 & -2 & 1 & \dots \\ 0 & 0 & 1 & -2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \quad (9)$$

Дифференцируя выражение (7) по z , получим:

$$\hat{z} = (E + \lambda D' D)^{-1} y, \quad (10)$$

где E - единичная матрица. Для более точной оценки в формулу (5) можно добавить весовые коэффициенты w_i , тогда:

$$S = \sum_i w_i (y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2, \quad (11)$$

где w_i - весовой коэффициент. Окончательное выражение для минимума примет вид:

$$\hat{z} = (W + \lambda D' D)^{-1} W y. \quad (12)$$

Данный алгоритм был реализован в среде MATLAB. Регулируя значения весовых коэффициентов w_i и λ , достигается аппроксимация базовой линии с наименьшей ошибкой. Результат определения базовой линии для модельного сигнала показан на рис. 2, а на рис. 3 - для реального сигнала.

3.2 Метод медианной фильтрации

Метод медианной фильтрации является одним из самых простых методов определения базовой линии. Для данного метода задается ширина окна фильтра в отсчетах. Формируется массив, который состоит из значений отсчетов сигнала внутри окна фильтра. Эти значения сортируются по возрастанию и на выход фильтра поступает элемент, который находится

посередине отсортированного массива. В случае четного количества элементов внутри окна выбирается среднее значение двух отсчетов в середине упорядоченного массива.

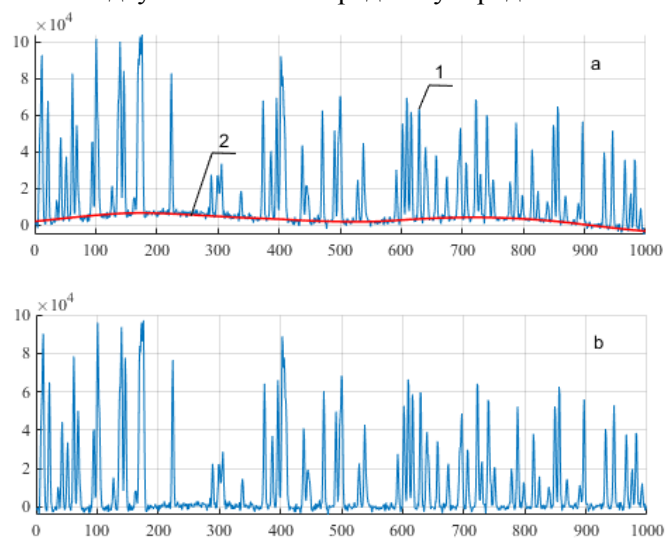


Рисунок 2. Модельный сигнал с базовой линией, определенной по методу АНК: а) График 1 - модельный сигнал, График 2 – расчетная базовая линия; б) график модельного сигнала после удаления базовой линии.

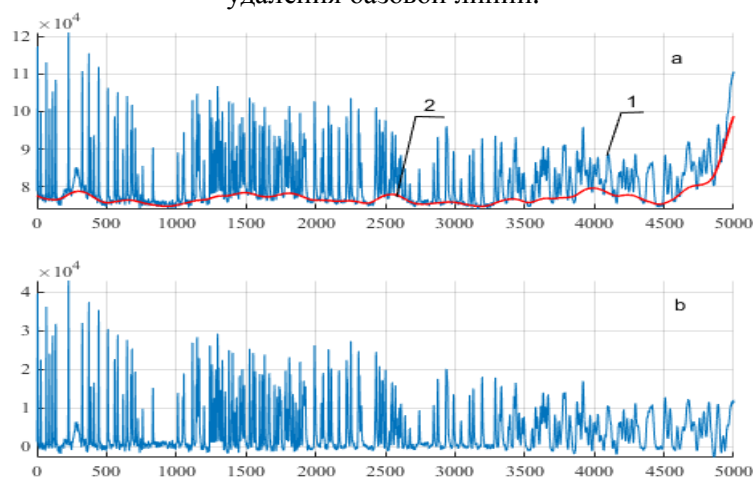


Рисунок 3. Реальный сигнал с базовой линией, определенной по методу АНК: а) График 1 - модельный сигнал, График 2 – расчетная базовая линия; б) график модельного сигнала после удаления базовой линии.

Была определена оптимальная ширина окна, при которой минимизируется среднеквадратичное отклонение расчетной базовой линии от заданной. Результаты для оптимального окна приведены на рис. 4 для модельного сигнала и на рис. 5 - для реального сигнала.

3.3 Метод на основе фильтра Савицкого-Голея

В последнее время для фильтрации данных и определения базовой линии стали часто использоваться фильтры Савицкого-Голея. В данной фильтрации в окрестности каждого отсчета сигнала строится аппроксимирующий полином n -го порядка по методу наименьших квадратов. Значение сигнала заменяется значением полинома в этой точке. Данный метод был реализован в среде MATLAB.

Для использования фильтра Савицкого-Голея, нужно задать степень полинома и ширину окна. Фильтр Савицкого-Голея выбрали нулевого порядка. Такой фильтр является модернизированным скользящим средним и не требует дополнительных параметров. Остается

выбор окна фильтра. В методе выбрано наименьшее окно из 3 отсчетов, которое увеличивается итеративно пока не будет достигнут критерий останова. В итоге получается полностью автоматический метод, подходящий для любого сигнала.

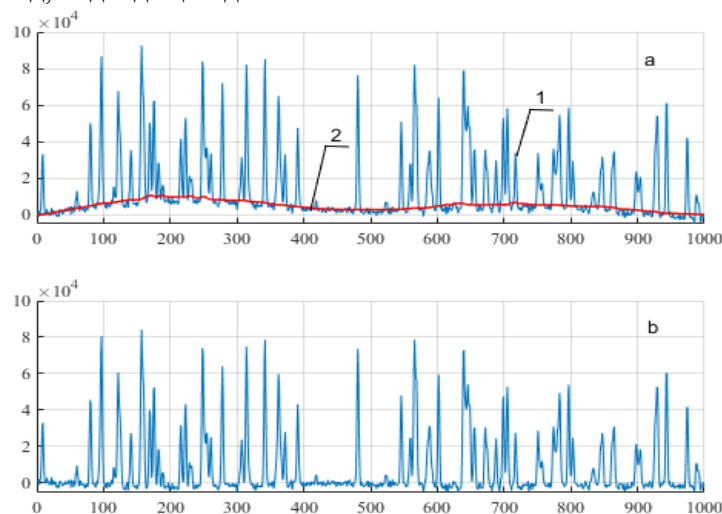


Рисунок 4. Модельный сигнал с базовой линией, определенной по методу медианного фильтра: а) График 1 – модельный сигнал, График 2 – расчетная базовая линия; б) график модельного сигнала после удаления базовой линии.

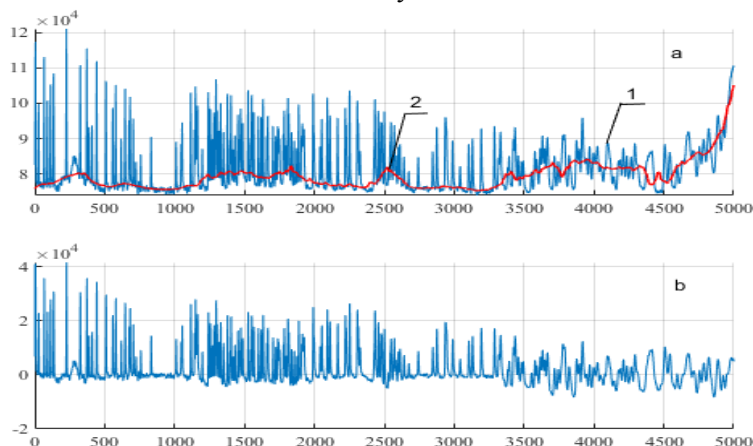


Рисунок 5. Реальный сигнал с базовой линией, определенной по методу медианного фильтра: а) График 1 – модельный сигнал, График 2 – расчетная базовая линия; б) график модельного сигнала после удаления базовой линии.

Начальными условиями является исходный сигнал y . Фильтруем y фильтром Савицкого-Голея с окном в 3 отсчета. В результате получаем некий y' . Рассчитываем площадь под y' , присваиваем в $y = y'$ и увеличиваем окно фильтра на единицу. Повторяем весь цикл до тех пор, пока площадь под кривой уменьшается. Результат для модельного сигнала приведен на рис. 6 для реального сигнала - на рис. 7.

4. Анализ полученных результатов

Для сравнения различных методов оценки базовой линии в модельных сигналах было выбрано два критерия. Первый критерий – это среднеквадратичное отклонение аппроксимированной базовой линии от истинной в модельном сигнале. Второй критерий – время вычисления оценки базовой линии. Данные по всем методам представлены в таблице 1.

АНК показывает наименьшую СКО по сравнению с другими методами. Метод медианной фильтрации имеет наименьшее время вычисления, но худшую СКО. Метод на основе

Савицкого-Голея не обладает ни низким СКО, ни временем вычисления, но зато является полностью автоматическим и не требует подбор оптимальных значений.

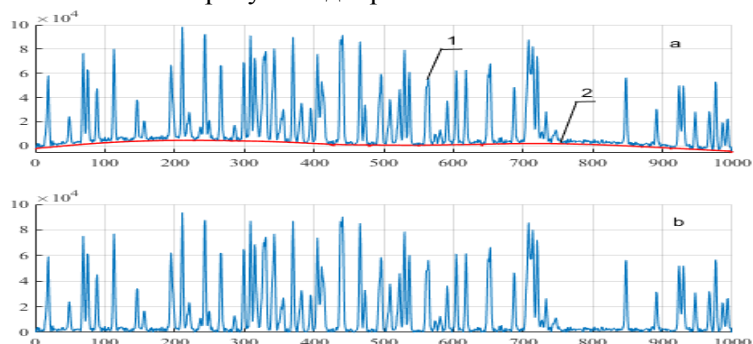


Рисунок 6. Модельный сигнал с базовой линией, определенной с помощью фильтра Савицкого-Голея: а) График 1 - модельный сигнал, График 2 – расчетная базовая линия; б) график модельного сигнала после удаления базовой линии.

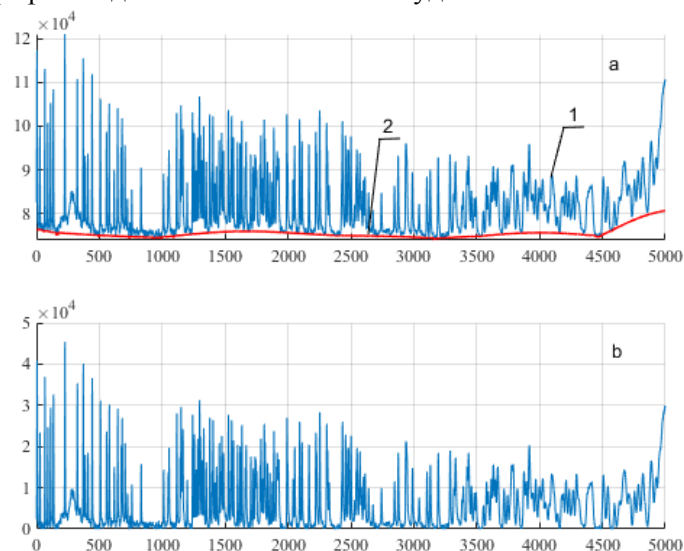


Рисунок 7. Реальный сигнал с базовой линией, определенной с помощью фильтра Савицкого-Голея: а) График 1 - модельный сигнал, График 2 – расчетная базовая линия; б) график модельного сигнала после удаления базовой линии.

Таблица 1. Среднеквадратичное отклонение и время вычисления базовой линии для различных методов.

Метод	СКО (отн. ед.)	Время вычисления (с)
АНК	$2.86 \cdot 10^5$	0.0087
Медианная фильтрация	$3.45 \cdot 10^6$	$5.6553 \cdot 10^{-4}$
На основе фильтра Савицкого-Голея	$3.28 \cdot 10^6$	0.0159

Также было проведено сравнение методов на реальных сигналах рис.8. Качественно можно сказать, что в некоторых областях сигнала АНК демонстрируют меньше искажений. В интервале 3500-4500 метод на основе медианного фильтра переоценивает базовый уровень, в то время как метод на основе фильтра Савицкого-Голея недооценивает базовый уровень.

5. Заключение

В статье выполнено моделирование сигналов с известной базовой линией со свойствами реальных сигналов генетического анализатора. Модельные сигналы использованы для

сравнения эффективности способов аппроксимации базовой линии различными методами. Было произведено сравнение методов и найден оптимальный по заданному критерию. Наилучшим по критерию СКО оказался метод АНК. Он имеет на порядок меньшую СКО по сравнению с методом медианной фильтрации. Метод на основе фильтра Савицкого-Голея имеет примерно тоже СКО, что и для медианной фильтрации, зато этот метод является полностью автоматическим и может применяться для любой физической модели. Сравнение методов на реальных сигналах не удалось осуществить численно, но качественно было показано, что метод АНК наиболее точно повторяет базовую линию сигналов.

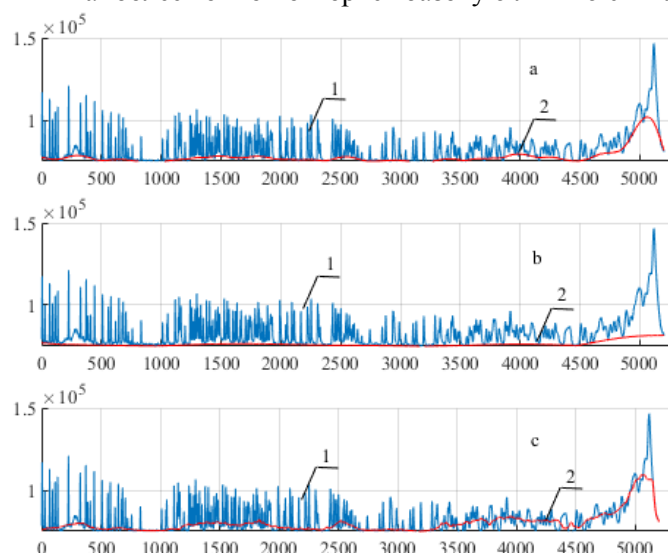


Рисунок 8. Базовая линия, определенная с помощью всех рассмотренных методов: а) График 1 – реальный сигнал, График 2 – базовая линия, рассчитанная по методу АНК; б) График 1 – реальный сигнал, График 2 – базовая линия, рассчитанная по методу на основе фильтра Савицкого-Голея; в) График 1 – реальный сигнал, График 2 – базовая линия, рассчитанная по методу на основе медианного фильтра.

6. Благодарности

Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований (грант 19-29-06036).

7. Литература

- [1] Белов, Ю.В. Особенности количественных измерений содержания нуклеиновых кислот методом полимерной цепной реакции в реальном времени / Ю.В. Белов, А.И. Петров, В.В. Иванов, И.У. Курочкин // Научное приборостроение. – 2011. – Т. 21, № 1. – С. 44-49.
- [2] Алексеев, Я.И. Приборы для диагностики биологических объектов на основе метода полимерной цепной реакции в реальном времени (ПЦР-РВ) / Я.И. Алексеев, Ю.В. Белов, Д.А. Варламов, С.В. Коновалова, В.Е. Курочкин, Н.Ф. Малышкин, А.И. Петров, А.О. Петряков, Д.А. Румянцев, С.Ю. Скоблилов, В.Н. Соколов, В. Н. Фесенко, А.В. Чернышев // Научное приборостроение. – 2006. – Т. 16, № 3. – С. 132-136.
- [3] Антонова, О.С. Полимерная цепная реакция: Приборная и методическая реализация. Обзор аналитических характеристик / О.С. Антонова, Г.Е. Рудницкая, А.Н. Тупик, А.Л. Буляница, А.А. Евстапов, В.Е. Курочкин // Научное приборостроение. – 2011. – Т. 21, № 4. – С. 5-21.
- [4] Давыдов, В.В. Многофункциональный малогабаритный ядерно-магнитный спектрометр / В.В. Давыдов, Н.С. Мязин // Измерительная техника. – 2017. – № 2. – С. 58-62.
- [5] Myazin, N.S. Features of formation of structure of a nuclear magnetic resonance signal in weak magnetic field / N.S. Myazin, V.V. Davydov // Journal Physics: Conference Series. – 2018. – 1135(1). – P. 012061.

- [6] Davydov, V.V. On the sensitivity of running-fluid NMR magnetometer / V.V. Davydov, V.I. Dudkin, A.A. Petrov, N.S. Myazin // *Technical Physics Letters*. – 2016. – Vol. 42(7). – P. 692-696.
- [7] Nepomnyashchaya, E.K. Spectroscopic techniques to study the immune response in human saliva / E.K. Nepomnyashchaya, E.A. Savchenko, E.N. Velichko, E.T. Bogomaz // *Journal of Physics: Conference Series*. – 2018. – Vol. 956(1). – P. 012009.
- [8] Baranov, M.A. Methods of non-destructive testing in studies of self-organization processes in protein films / M.A. Baranov, E.N. Velichko, E.T. Aksenov // *Journal of Physics: Conference Series*. – 2017. – Vol. 917(1). – P. 062059.
- [9] Davydov, R.V. Monitoring of flowing media state by refraction phenomenon / R.V. Davydov, V.I. Antonov, A.V. Moroz // *Proceedings of the IEEE International Conference on Electrical Engineering and Photonics*. – 2018. – Vol. 8564378. – P. 236-239.
- [10] Davydov, R.V. Equation of state for computer simulation of metal ablation by femtosecond laser pulses / R.V. Davydov, V. I. Antonov // *Journal of Physics: Conference Series*. – 2017. – Vol. 929(1). – P. 012040.
- [11] Белов, Ю.В. Коррекция базовой линии сигнала флуоресцентного детектора генетического анализатора / Ю.В. Белов, И.А. Леонтьев, А.И. Петров, В.Е. Курочкин // *Научное приборостроение*. – 2013. – Т. 23, № 2. – С. 9-13.
- [12] Алексеев, Я. И. Исследование погрешности оцифровки пиков генетического анализатора / Я.И. Алексеев, Ю.В. Белов, О.П. Малюченко // *Научное приборостроение*. – 2014. – Т. 24, № 2. – С. 79-85.
- [13] Eilers, P.H.C. Baseline Correction with Asymmetric Least Squares Smoothing / P.H.C. Eilers, H.F.M. Boelens – London: Taylor & Francis Ltd., 2005. – 142 p.
- [14] Schulze, H. A Small-Window Moving Average-Based Fully Automated Baseline Estimation Method for Raman Spectra / H. Schulze, R. Foist, K. Okuda, A. Ivanov, R. Turner // *Applied spectroscopy*. – 2012. – Vol. 66. – P. 757-764.
- [15] Whittaker, E.T. On a new method of graduation / E.T. Whittaker // *Proceedings of the Edinburgh Mathematical Society*. – 1922. – Vol. 41. – P. 63-73.

Comparison of methods for baseline determining of fluorescent detector signals of genetic analyzer

Y.V. Batov¹, D.A. Puzko¹, A.I. Petrov², V.V. Davydov^{1,3}, V.Yu. Rud³

¹Higher School of Applied Physics and Space Technologies, Peter the Great Saint Petersburg Polytechnic University, Polytechnicheskaya str. 29, Saint Petersburg, Russia, 195251

²Institute of Analytical Instrumentation of the Russian Academy of Sciences, Riga ave. 26, St. Petersburg, Russia, 190103

³All-Russian Research Institute of Phytopathology, Institut str. 5, B. Vyazemy, Moscow Region, Russia, 143050

Abstract. In article of the methods for baseline determining of signals from a fluorescent detector of a genetic analyzer are discussed. The simulated for signal which has characteristics identical to the characteristics of a real signal is performed. The comparison of the methods to the criterion of mean square error and calculation speed for model signals is done. A qualitative analysis of the effectiveness of methods on real signals is carried out. The theoretical calculations by these methods and research results are presented.