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

Y V Batov¹, D A Puzko¹, A I Petrov², V V Davydov¹,³, V Yu Rud³

¹Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russia, 195251
²Institute of Analytical Instrumentation of the Russian Academy of Sciences, St. Petersburg, Russia, 190103
³All-Russian Research Institute of Phytopathology, 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.

1. Introduction
One of the methods used in genetic analysis is Sanger DNA sequencing [1-3]. This method has several advantages in terms of informativeness of research in comparison with other methods, for example, optical spectroscopy [4-9] or nuclear magnetic spectroscopy [10-15]. The latter have a high accuracy of measurement, but the information contained in the spectrum is not enough for genetic analysis with a clear result.

In the process of sequencing, DNA is divided into fragments in increments of 1 nucleotide. At the end of each fragment, a fluorescent dye is placed depending on the last nucleotide in the fragment (a total of 4 dyes are used, one for each nucleotide). DNA fragments are separated in the capillary under the influence of an electric field [1-3]. At the end of the capillary is a laser that excites a fluorescent dye; this excitation is detected by a special detector. This is a significant difference in the implementation of this method from the classical ones used for research in optical devices [16-24].

4 channels are used for sequencing. Data from the detector output is digitized in the form of 4 digital sequences. The horizontal axis shows the numbers of signal samples, and the vertical axis shows the light intensity in relative units.

Processing detector signals is a non-trivial task. Due to the hardware properties, the presence of noise and other factors affecting laser detection, noise appears, a baseline and peak merging. Usually the first task is to remove the baseline. There are many methods for determining the baseline. One of the simplest is median filtering. This method is used in a genetic analyzer developed by IAI RAS together with Synthol CJSC [25]. However, when using this method, significant errors arise that can distort the original data.

The paper considers alternative methods for eliminating the baseline, which have significantly less errors in comparison with median filtering. These methods are applied to the model signal and to real data. A comparison is made according to the criterion of the minimum mean squared error (MMSE) of the proposed methods.
2. Modeling the ideal signal from the detector output of the genetic analyzer

To calculate the standard error of the found baseline, it is necessary to know the specified baseline. The signal was simulated for this. Based on the analysis of several methods [26-29] and considering the features identified by us in the experimental data, the following methodology was chosen.

An array of length 600 is generated, each element of which is a random number from 1 to 4 (each corresponds to its own nucleotide). The masses of nucleotides and dyes are specified [26]. Fragments are created from nucleotides in steps of 1, that is, the first fragment consists of the first nucleotide, the second of the first two, etc. The dye mass is added to the calculated mass of each fragment depending on the type of the last nucleotide in the fragment.

\[ m_i = \sum_{j=1}^{i} m_{nj} + m_{di}, \]  

where \( m_i \) is the mass of the \( i \)th fragment, \( m_{nj} \) is the mass of the \( j \)th nucleotide, \( m_{di} \) is the mass of the dye corresponding to the \( i \)th nucleotide.

Considering the movement of fragments in the capillary uniformly accelerated and knowing the mass, we find the time for which each fragment arrives. At these points, peaks form in one of the four signals. The peak type is approximately considered Gaussian. The width of the Gaussian peak increased with increasing time; this effect is observed in real signals. It was also noted that in real signals the peak amplitude decreases with increasing time, but there are also peaks whose amplitudes change randomly. In the model signal, the amplitude \( A \) is generated as follows:

\[ A = \alpha t + x, \]  

where \( \alpha \) characterizes the general linear dependence, \( t \) is the time, \( x \) is the value from a uniform distribution from 0 to 1.

Thus, the overall increase in amplitude remains linear, but each specific value was a random variable. The final equation of the ideal model signal \( s \) can be represented as follows:

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

where \( \sigma_i \) is the width of the Gaussian peak (in terms of the normal distribution (ND) is the variance), \( m_i \) is the mass of the \( i \)th fragment (in terms of the NR is the mean (location parameter)).

Noise and baseline were added to the received model signal (3). Baseline \( B(t) \) was defined as:

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

The noise was superimposed on the baseline by adding to each value of the baseline value from a random, uniformly distributed value from 0 to 1. Figure 1 shows the type of model signal. MATLAB is selected as a modeling environment.

3. Baseline determination methods

In modern genetic analysis, various methods are used to remove the baseline. Frequently used methods are the method of asymmetric least squares (ALS) [30] and the method based on the Savitsky-Golay filter [31]. These methods were compared with the median filtering method and the optimal one was found by the MSE criterion and calculation time.

3.1 Method of asymmetric least squares

The asymmetric least squares method is based on the Whittaker method with weighting coefficients [32]. Whittaker smoothing is defined by the following:

\[ S = \sum_i (y_i - z_i)^2 + \lambda \sum_i (A^2z_i)^2, \]  

\[ \Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}) = z_i - 2z_{i-1} + z_{i-2} \]  

(6) 

where \( y_i \) - samples of the real baseline, \( z_i \) - samples of the approximated baseline, \( \lambda \) - weight coefficient. The first term in (5) is the standard deviation of the approximated baseline from the real one; the second term determines smoothing. Find the minimum of S. Rewrite the expression in matrix form:

\[
S = (y - z)'(y - z) + \lambda z'D'Dz, \\
y = (y_1, ..., y_n)', z = (z_1, ..., z_n)', \\
D = \begin{bmatrix}
1 & -2 & 1 & 0 & \ldots \\
0 & 1 & -2 & 1 & \ldots \\
0 & 0 & 1 & -2 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\]  

(9) 

Figure 1. Modeling the signal from the detector output. Graph a) large scale. Graph b) signal fragment of a). 

Differentiating expression (7) with respect to \( z \), we obtain:

\[
\hat{z} = (E + \lambda D'D)^{-1}y, \\
E \text{ is the unit matrix. For a more accurate estimate, we can add weight coefficients } w_i \text{ to formula (5), then:}
\]

\[
S = \sum \omega_i(y_i - z_i)^2 + \lambda \sum (\Delta^2 z_i)^2, 
\]

(11) 

where \( \omega_i \) - weight coefficient. The final expression for the minimum takes the form:

\[
\hat{z} = (W + \lambda D'D)^{-1}Wy. 
\]

This algorithm was implemented in MATLAB. By adjusting the values of the weighting coefficients \( w_i \) and \( \lambda \), the approximation of the baseline with the least error is achieved. The result of determining the baseline for the model signal is shown in figure. 2 and in figure. 3 - for a real signal.

3.2 Median filtering method 
The median filtering method is one of the simplest methods for determining the baseline. For this method, the filter window width in samples is set. An array is formed, which consists of signal samples within the filter window. These values are sorted in ascending order and the element that is in
the middle of the sorted array arrives at the filter output. In the case of an even number of elements inside the window, the average value of two samples in the middle of the ordered array is selected.

The optimal window width was determined at which the standard deviation of the calculated baseline from the specified one is minimized. The results for the optimal window are shown in figure 4 for the model signal and in figure 5 for a real signal.

![Figure 2](image1.png)

**Figure 2.** Model signal with a baseline determined by the ALS method: a) Graph 1 - model signal, Graph 2 - calculated baseline; b) a graph of the model signal after removing the baseline.

![Figure 3](image2.png)

**Figure 3.** A real signal with a baseline determined by the ALS method: a) Graph 1 - model signal, Graph 2 - calculated baseline; b) a graph of the model signal after removing the baseline.

### 3.3 Method based on the Savitsky-Golay filter

Recently, Savitsky-Golay filters have often been used to filter data and determine the baseline. In this filtering, in the vicinity of each signal sample, an approximating polynomial of the nth order is constructed using the least squares method. The signal value is replaced by the polynomial value at this point. This method was implemented in MATLAB.

To use the Savitsky-Golay filter, you need to set the degree of the polynomial and the width of the window. The Savitsky-Golay filter chose zero order. Such a filter is a modernized moving average and
does not require additional parameters. The choice of filter window remains. In the method, the smallest window of 3 samples was selected, which increases iteratively until the stop criterion is reached. The result is a fully automatic method suitable for any signal.

The initial conditions are the original signal $y$. Denote by $y'$ the result of filtering $y$ by the Savitsky-Golay filter with a window of 3 reports. The area under $y'$ is calculated, $y = y'$ is assigned and the filter window is increased by one. The algorithm is repeated until the area under $y'$ decreases by an amount that is less than $\varepsilon$. The result of finding the baseline for the model signal is shown in figure 6 for a real signal - in figure 7.

4. Analysis of the results

Two criteria were chosen to compare different methods for estimating the baseline in model signals. The first criterion is the standard deviation of the approximated baseline from the true in the model signal. The second criterion is the time for calculating the baseline estimate. Data for all methods are presented in table 1.

Figure 4. Model signal with a baseline determined by the median filtering method: a) Graph 1 - model signal, Graph 2 - calculated baseline; b) a graph of the model signal after removing the baseline.

Figure 5. A real signal with a baseline determined by the median filtering method: a) Graph 1 - model signal, Graph 2 - calculated baseline; b) a graph of the model signal after removing the baseline.
Figure 6. Model signal with a baseline determined using the Savitsky-Golay filter: a) Graph 1 - model signal, Graph 2 - calculated baseline; b) a graph of the model signal after removing the baseline.

Figure 7. Real signal with a baseline determined using the Savitsky-Golay filter: a) Graph 1 - model signal, Graph 2 - calculated baseline; b) a graph of the model signal after removing the baseline.

Table 1. Minimum mean squared error and baseline calculation time for various methods

| Method                      | MMSE (arb. unit) | Calculation time (s) |
|-----------------------------|------------------|----------------------|
| ALS                         | $2.86 \times 10^5$ | 0.0087               |
| Median Filtering            | $3.45 \times 10^6$ | $5.6553 \times 10^{-4}$ |
| Method based on the Savitsky-Golay filter | $3.28 \times 10^6$ | 0.0159               |

ALS shows the lowest standard deviation in comparison with other methods. The median filtering method has the shortest calculation time, but the worst standard deviation. The Savitsky-Golay method
does not have either a low standard deviation or calculation time, but it is completely automatic and does not require the selection of optimal values.

A comparison was also made of methods using real signals in figure 8. Qualitatively, it can be said that in some areas of the signal, ALSs exhibit less distortion. In the interval 3500-4500, the method based on the median filter overestimates the base level, while the method based on the Savitsky-Golay filter underestimates the base level.

Figure 8. The baseline determined using all the methods considered: a) Graph 1 - the real signal, Graph 2 - the baseline calculated by the ANC method; b) Graph 1 - real signal, Graph 2 - baseline calculated by the method based on the Savitsky-Golei filter; c) Graph 1 - real signal, Graph 2 - baseline calculated by the method based on the median filter;

5. Conclusion
The article simulates signals with a known baseline with the properties of real signals from a genetic analyzer. Model signals are used to compare the effectiveness of the methods for approximating the baseline by various methods. A comparison of the methods was made and the optimal one was found by the given criterion. The best according to MSE criterion was the method of ALS. It has an order of magnitude lower MSE compared to the median filtering method. The method based on the Savitsky-Golay filter has approximately the same MSE as for median filtering, but this method is fully automatic and can be used for any physical model. A comparison of methods using real signals could not be carried out numerically, but it was qualitatively shown that the method of ALS most accurately repeats the signal baseline.

6. References
[1] Belov Yu V, Petrov A I, Ivanov V V and Kurochkin V E 2011 Scientific instrumentation 21(1) 44-49.
[2] Alekseev Ya I, Belov Yu V, Malyuchenko O P, Monakhova Yu A, Natyrova A N, Orekhov V A, Konovalov S V, Kurochkin V E and Petrov A I 2012 Scientific instrumentation 22(4) 86-92.
[3] Antonova O S, Rudnitskaya G E, Tupik A N, Bylianitsa A L and Evstrapov V E 2011 Scientific instrumentation 21(4) 5-21.
[4] Skidanov R V, Blank V A and Morozov A A Study of an imaging spectrometer based on a diffraction lens Computer Optics 39(2) 218-223 DOI: 10.18287/0134-2452-2015-39-2-218-223.
Kozlova E S and Kotlyar V V 2014 Specified parameters of sellmeyer model for silica glass Computer Optics 38(1) 51-56.

Kuzmin M S, Davydov V V and Rogov S A 2019 On the use of a multi-raster input of one-dimensional signals in two-dimensional optical correlators Computer Optics 43(3) 391-396. DOI: 10.18287/2412-6179-2019-43-3-391-396.

Vasilichenko V A, Burkovskiy V L and Danilov A D 2019 Algorithmization of the process of recognition of states of living objects based on special X-ray images Computer Optics 43(2) 296-303. DOI: 10.18287/2412-6179-2017-43-2-296-303.

Perminov N S, Petrovvin K V, Gerasimov K I, Kirillov R R, Sherstvukov O N and Moiseev S A 2018 Spectroscopy of cascade multiresonator quantum memory Computer Optics 42(4) 614-619. DOI: 10.18287/2412-6179-2018-42-4-614-619.

Morozov A N, Kochikov I V, Novgorodskaya A V, Sologub A A and Fufurin I L 2015 Statistical estimation of the probability of the correct substance detection in FTIR spectroscopy Computer Optics 39(4) 614-21. DOI: 10.18287/0134-2452-2015-39-4-614-621.

Davydov V V, Cheremiskina A V, Velichko E N and Karseev A Yu 2014 Journal of Physics: Conference Series 541(1) 012006.

Baranov M A, Dudina A I and Nepomnyashchaya E K 2019 Journal of Physics: Conference Series 1226(1) 012005.

Nepomnyashchaya E, Velichko E, Aksenov E and Bogomaz T 2018 Biophotonics: Photonic Solutions for Better Health Care VI 10685 106852F.

Savchenko E and Velichko E 2018 Proc. Int. Conf. on Optical and Nano-Technologies for Biology and Medicine (Saratov) 11065, 110651U.

Davydov R V, Antonov V I and Moroz A V 2018 Proceedings of the IEEE International Conference on Electrical Engineering and Photonics (Saint-Petersburg) 8564378 236-239.

Davydov R V, Mazing M S, Yushkova V V and Stimirianov A V 2019 Journal of Physics: Conference Series 1410(1) 012067.

Davydov R, Antonov V, Makeev S, Batov Y, Dudkin V and Myazin N 2019 E3S Web of Conferences 140 02001.

Grevtseva A S, Smirnov K J, Grebenikova N M and Glinushkin A P 2019 Journal of Physics: Conference Series 1368(2) 02072.

Smirnov K J, Medzakovskiy V I, Vysoczky M G and Glagolev S F 2017 Journal of Physics: Conference Series 919(6) 062019.

Davydov R V and Antonov V I 2017 Journal of Physics: Conference Series 929(1) 012040.

Belov Yu V, Leont’ev I A, Petrov A I and Kurochkin V E 2013 Scientific instrumentation 23(2) 9-13.

Alekseev Y I, Belov Yu V and Maluchenko O P 2014 Scientific instrumentation 24(2) 79-85.

Chernov V M 2015 Quasiparallel algorithm for error-free convolution computation using reduced Mersenne–Lucas codes Computer Optics 39(2) 241-248. DOI: 10.18287/0134-2452-2015-39-2-241-248.

Filimonov P A, Belov M L, Fedotov Y V, Ivanov S E and Gorodnichev V A 2018 An algorithm for segmentation of aerosol inhomogeneities Computer Optics 42(6) 1062-1067. DOI: 10.18287/2412-6179-2018-42-6-1062-1067.

Denisova A Y and Sergeyev V V 2016 Algorithms for calculating multichannel image histogram using hierarchical data structures Computer Optics 40(4) 535-542. DOI: 10.18287/2412-6179-2016-40-4-535-542.
[30] Paul H C 2005 Boelens (London: Taylor & Francis Ltd.).
[31] Schulze H, Foist R, Okuda K, Ivanov A and Turner R 2012 Applied spectroscopy 66 757-764.
[32] Whittaker E T 1922 Proceedings of the Edinburgh Mathematical Society 41 63-73.