Using wavelet transform to detect peaks in PCR signals

Y V Batov¹, D A Puzko¹, A I Petrov², V V Davydov¹-³, A P Glinushkin³

¹Peter the Great Saint Petersburg Polytechnic University, St. Petersburg 195251, Russia
²Institute of Analytical Instrumentation of the Russian Academy of Sciences, St. Petersburg 190103, Russia
³All Russian Research Institute of Phytopathology, Moscow Region 143050, Russia

Abstract. The article of the feasibility of determining the nucleotide sequence of sequencing to Sanger is justified. The signals from a sample obtained using a DNA analyzer in which there is a nucleotide sequence are reviewed. A method for determining the sequence based on a continuous wavelet transform is proposed. The results of experimental investigations are presented.

1. Introduction
At present, many methods have been developed and successfully applied to study various environments [1–9]. Many of these methods, especially optical and nuclear magnetic, allow us to determine the structure and composition of the environment [10–18]. The chemical bonds between the molecules are determined when using high-resolution spectrometers using these methods [13, 15, 17, 19–26]. In some cases, this allows us to solve some problems related to biology, virology, and forensics [19, 20, 25–27]. The development of science has shown that even the use of new methods for processing the optical signal [27–38] in these devices does not allow obtaining the necessary data for making a reliable decision. Therefore, a separate class of optical devices (DNA analyzers) was created to solve these problems [39–41]. The various methods are used to obtain the nucleotide sequence in these devices [39–43]. The most widely used method is Sanger sequencing.

Dye sequencing is currently the basis of automated sequencing, due to its feasibility and high speed. The disadvantage of this method is the difference in the combination of dye-labeled chain terminators with a DNA fragment, which leads to a change in the peak shape in the chromatogram of the electronic DNA sequence after capillary electrophoresis.

The sequencers of the DNA are detected and recorded the dye fluorescence and output data as a sequence of peaks using capillary electrophoresis. Several software packages are implemented that can evaluate the reliability of the peak and remove low-quality ones. The during a research of the DNA on the viruses presence of viruses in it, the number of peaks increases (their amplitude decreases). This requires the development of new information processing methods. The article considers one of the options. We have proposed a method for identifying peaks based on the wavelet transform, which can improve the accuracy of detection of reliable peaks.

2. Wavelet transform
Wavelet transform is an integral transform, which is a convolution of a mother wavelet with a signal. The wavelet transform translates the signal from a temporal representation into a frequency-temporal one.

The wavelet transform for a continuous signal relative to the mother wavelet is defined as follows:

\[ Wf(u,s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-u}{s}\right)dt, \]  

(1)
where the parameter $u \in R$ corresponds to the time shift, and is called the position parameter, the parameter $s > 0$ sets the scaling and is called the stretching parameter.

Like a windowed Fourier transform, a wavelet transform can measure the time evolution of frequency transients. This requires using a complex analytic wavelet, which can separate amplitude and phase components. In contrast, real wavelets are often used to detect sharp signal transitions [44-46]. Since the purpose of using the wavelet transform is to detect peaks in the signal, a real wavelet was used.

The based on the definition of CWT (Continuous Wavelet Transform), the wavelet coefficients are found from the internal product between the signal $x(t)$ and the wavelets $\psi_{s,u}(t)$. The wavelet coefficients indicate the similarity of the signal by the transformation index $u$ and the wavelet in the scale $s$. To detect peaks, the goal is to map the various shapes (heights and widths) of the peaks by scaling the mother wavelet $\psi(t)$.

It can be shown that the local smoothness of the signal is characterized by a decrease in the amplitude of the wavelet transform with a decrease in scale. Features and differences are highlighted by the study of local maximums of the wavelet transform at small scales [44-46].

It was shown in [41, 42] that the peaks in PCR signals can be approximately considered a Gaussian function. The mother wavelet is selected so that its main characteristics are similar to the desired peak in the signal [47]. The Mexican hat wavelet is most suitable for this study because of the similarity of its shape to the peak.

The Mexican hat wavelet is a Ricker wavelet in mathematics and numerical analysis and is defined as:

$$\psi(t) = \frac{1}{\sqrt{2\pi}}(1-t^2)e^{-\frac{t^2}{2}}$$

3. Algorithm implementation and signal processing results

It follows from the Hwang – Mallat theorem that the function $f$ can be singular (not Lipschitz 1) at a point $\nu$, only if there is a sequence of wavelet maxima points $(u_p, s_p), p \in \mathbb{N}$ that converges toward $\nu$ at fine scales [44]. Thus, the essence of the method is to determine the features of the signal, that is, in our case, peaks from the lines of the maxima on the signal scalogram. The following is an algorithm for determining peaks using a continuous wavelet transform.

3.1. Modeling input signal

The signal modeling was performed according to [41]. The sampling frequency was chosen in such a way that about 10 samples fall on one peak, which corresponds to real data. Denote the model signal as:

$$x[k], \quad k = 0,1,2,\ldots,N,$$

where $N$ is the length of signal.

3.2. Continuous wavelet transform

We denote the value $s_j$ of the $j$-th scaling factor $j = 0,1,2,\ldots,J$. Then the wavelet transform of $x[k]$ with the mother wavelet $\psi[n]$ can be written as:

$$X_{j,k} = (X)_{j,k} = \sum_{n=-\infty}^{\infty} x[n]\psi_{s_j,n}[n], \quad j = 0,1,\ldots,J, \quad k = 0,1,\ldots,N,$$

where $X$ is the matrix of wavelet transform coefficients. Accordingly, the wavelet coefficients for the scale $s_j$ are denoted as $X_{j,:}$, the $j$-th row of the matrix.

3.3 Maxima detection

From the obtained matrix $X$, we determine all peak maxima for each scale, that is, it is necessary to satisfy the condition:

$$P_j = \{k \mid X_{j,k} > \gamma, \quad X_{j,k} > X_{j,k-1}, \quad X_{j,k} > X_{j,k+1}\}, \quad j = 0,1,2,\ldots,J,$$
where $\gamma$ is the threshold for deciding that this maximum is a peak. All maxima below the threshold are not considered. Therefore, $P_j$ contains all the maxima for the scale $s_j$, which are larger than $\gamma$. Sort the elements $P_j$ in ascending order:

$$0 \leq p_{j,k_1} < p_{j,k_2} \leq N, \quad 1 < k_1 < k_2 < m_j,$$

$p_{j,i}$ is the column index of the $i$-th peak maxima at row $s_j$ in matrix $X$, and $m_j$ is the total number of elements in $P_j$.

3.4. Identify ridges
All maxima from the matrix $P$ are studied from the largest scale to the smallest scale. The algorithm for identifying ridges is as follows:
1. Starts at largest scale, $j=J$. Every peak maxima at the largest scale $s_j$ starts a new ridge.
2. Go to the next scale by updating scale index: $j = j - 1$. If the peak maxima is close to a peak maxima at scale $s_{j+1}$, it will be connected to the existing ridge containing that peak maxima at scale $s_{j+1}$. If it is not close to any peak maxima at scale $s_{j+1}$, then a new ridge $s_j$ is created starting at this peak maxima.
3. Repeat Step 2 until it finishes processing peak maxima at the smallest scale.
Let us denote for the set of all found ridges $R_i = \{r_{x_{i,j}}, r_{y_{i,j}}\}, \quad i = 1,2,3,\ldots,N_R,\quad j = 1,2,3,\ldots,L_i$, where $N_R$ is the total number of detected ridges, $L_i$ is the number of maxima along the ridge, $r_{x_{i,j}}$ and $r_{y_{i,j}}$ denote the index of the row and column of the $j$-th maximum along $R_i$.
4. Lines of maxima whose length is less than $l_{max}$ are considered caused by noise and removed
5. Identify peak position
The peak position is determined $r_{y_i}$ for minimum scale in $R_i$.

The illustrates for peak definition are presented in figure.1.

![Figure 1](image-url)  
Figure 1 (a, b). The peak detection process: (a) – model signal; (b) – CWT with identified ridges after removing short ridges.
The effectiveness of our method was tested using the BLAST program on the Synthetic construct BigDye Terminator Cycle Sequencing Standard sequence. The first half of the available standard sequences were taken, in which there are no serious signal distortions and the entire sequence. The results obtained for one of the studied sequences are shown in table 1.

| Method               | Max score | Total score | Query cover | Percent identity |
|----------------------|-----------|-------------|-------------|------------------|
| First half of the signal | 649       | 649         | 100%        | 100%             |
| Whole signal         | 933       | 933         | 91%         | 96.08%           |

On average, about 99% of the query cover was achieved for the first half of the real signal. The results for the entire signal are highly dependent on the quality of that signal.

4. Conclusion
For real signals that do not have large distortions we were able to achieve a peak determination reliability of 0.99. The research shows that for the “good” part of the real signal, in which distortions are not yet very pronounced, this method detects on average more luminescent peaks than other methods. For better detection of peaks along the entire length of the signal, it is necessary to configure the initial parameters for constructing lines on the scaglogram. The further development of our work will be the automation of parameter settings for determining peaks.

References
[1] Rumyantsev N, Bondareva O, Makeev S and Krasnoshekov V 2019 IOP Conference Series: Earth and Environmental Science 390(1) 012037
[2] Kozar M, Sabliy L, Korencuk M, Korshunov A and Kosolapov V 2019 IOP Conference Series: Earth and Environmental Science 390(1) 012002
[3] Artem’ev K, Kolik L, Podkovyrov L, Sevostyanov S, Kosolapov V, Meshalkin V and Diuldin M 2019 IOP Conference Series: Earth and Environmental Science 390(1) 012039
[4] Shafeev G, Barmina E, Valiullin L, Simakin A, Ovsyankina A, Demin D, Kosolapov V, Korshunov A and Roman Denisov 2019 IOP Conference Series: Earth and Environmental Science 390(1) 012041
[5] Krasnova E D, Matorin D N, Belevich T A, Voronov D A and Patsaeva S V 2018 Journal of Oceanology and Limnology 36(6) 1962-1977
[6] Kosolapov V, Rud’ V, Korshunov A, Savchenko I, Switala F and Hogland W 2019 IOP Conference Series: Earth and Environmental Science 390(1) 012010
[7] Davydov R, Antonov V, Makeev S, Batov Y, Dudkin V and Myazin N 2019 E3S Web Conference 140 02001
[8] Semenov V V, Nikiforov N F, Ermak S V and Davydov V V 1991 Soviet Journal of Communications Technology and Electronics 36 59 – 63
[9] Davydov R, Antonov V and Kalinin N 2015 Journal of Physics: Conference Series 643(1) 012107
[10] Karseev A Yu, Vologdin V A and Davydov V V 2015 Journal of Physics: Conference Series 643(1) 012108
[11] Grebenikova N M, Myazin N S, Rud’ V Yu and Davydov R V 2018 Proceedings of the 2018 IEEE International Conference on Electrical Engineering and Photonics, EExPoltech 2018 (Saint-Petersburg) 8564409 p. 295-297
[12] Davydov R, Antonov V and Moroz A 2019 Proceedings of the 2019 IEEE International Conference on Electrical Engineering and Photonics, EExPoltech 2019 (Saint-Petersburg) 8906791 p. 42-45
[13] Davydov V V, Myazin N S and Davydo C T I 2017 Russian Journal of Nondestructive Testing 53(7) 520-529
[14] Zhernovoi A I, Komlev A A and D’yachenko S V 2016 Technical Physics 61(2) 302-305
[15] Davydov V V and Myazin N S 2017 Measurement Techniques 60(5) 491-496
[16] Davydov V V, Dudkin V I and Karseev A Y 2015 Journal of Applied Spectroscopy 82(5) 794-800
[17] Davydov V V, Velitchko E N, Dudkin V I and Karseev A Y 2015 

Instruments and Experimental
Techniques 58(2) 234-238

[18] Davydov V V 1999 Russian Physics Journal 42(9) 822-825

[19] Gizatullin B, Gafurov M, Rodionov A, Mamin G, Mattea C, Stapf S and Orlinskii S 2018 Energy
 Fuels 32(11) 11261-11268

[20] Alexandrov A S, Ivanov A A, Archipov R V, Gafurov M R and Tagirov M S 2019 Magnetic
 Resonance in Solids 21(2) 19203

[21] Davydov V V, Dudkin V I and Karseev A Y 2015 Instruments and Experimental Techniques 58(6) 

787-793

[22] Davydov V V, Dudkin V I and Karseev A Y 2014 Measurement Techniques 57(8) 912-918

[23] D'yachenko S V and Zhernovoy A I 2016 Technical Physics 61(12) 1845–1847

[24] Diachenko S V, kondrashkova I S and Zhernovoy A I 2017 Technical physics 62(10) 1602–1604

[25] Davydov V, Cheremiskina A, Velitchko E and Karseev A 2014 Journal of Physics: Conference Series

541(1) 012006

[26] Davydov V V, Velitchko E N, Dudkin V I and Karseev A Y 2014 Measurement Techniques 57(6) 684- 

689

[27] Davydov R V, Antonov V I, Yushkova V V, Grebenikova N M and Dudkin V I 2019 Journal of

Physics: Conference Series 1236(1) 012079

[28] Grebenikova N M, Davydov R V and Rud V Yu 2019 Journal of Physics: Conference Series 1326(1) 

012012

[29] Yushkova V, Kostin G, Davydov R, Rud S, Dudkin V and Valiullin L 2019 IOP Conference Series: 

Earth and Environmental Science 390(1) 012016

[30] Logunov S E, Rud V Yu, Davydov R V, Moroz A V and Smirnov K J 2019 Journal of Physics: 

Conference Series 1326(1) 012024

[31] Valov A P, Davydov R V, Rud V Yu and Grevtseva A S 2019 Journal of Physics: Conference Series

1326(1) 012040

[32] Grevtseva A S, Davydov R V, Dudkin V I and Rud' V Yu 2019 Journal of Physics: Conference Series

1326(1) 012043

[33] Lukashev N A, Davydov R V, Glinushkin A P and Rud' V Yu 2019 Journal of Physics: Conference Series

1326(1) 012046

[34] Logunov S E, Davydov R V, Vysotsky M G, Dudkin V I and Rud' V Yu 2019 Journal of Physics: 

Conference Series 1368(1) 022056

[35] Davydov R V and Antonov V I 2016 Journal of Physics: Conference Series 769(1) 012060

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

[37] Davydov R V and Antonov V I 2018 Journal of Physics: Conference Series 1124(1) 081037

[38] Davydov V V, Dudkin V I, Petrov A A and Myazin N S 2016 Technical Physics Letters 42(7) 

692-696

[39] Alekseev J I, Belov Yu V, Varlamov D A, Konovalov C V, Kurochkin V E, Marakushin N F, Petrov 

A I, Petryakov A O, Rumyantsev D A, Skolobilov S Yu, Sokolov V N, Fresenko V N and Chernewshov 

A V 2006 Scientific instrument engineering 16(3) 132-136

[40] Antonova O S, Rudnizky G E, Tupik A N, Byliianiza A L, Evstrapov A A and Kurochkin V E 2011

Scientific instrument engineering 21(4) 5-21

[41] Belov Yu V, Petrov A I, Ivanov V V and Kurochkin V E 2011 Scientific instrumentation 21(1) 44- 

49

[42] Belov Yu V, Leont’ev I A, Petrov A I and Kurochkin V E 2013 Scientific instrumentation 23(2) 9– 

13

[43] Alekseev Y I, Belov Yu V and Maluchenko O P 2014 Scientific instrumentation 24(2) 79–85

[44] Mallat S 1999 A wavelet tour of signal processing (New York: Wiley–Interscience)

[45] Bozhokin S V, Lesova E M, Samoilov V O, et al. 2018 Hum Physiol. 44 32–40

[46] Bozhokin S V and Sokolov I M 2018 Tech. Phys. 63 1711–1717

[47] Du P, Kibbe W, Simon M. 2006 Bioinformatics 22(17) 2059–2065