14

Image processing algorithms in the DNA sequencer „Nanofor SPS“

© V.V. Manoilov, A.G. Borodinov, A.S. Saraev, A.I. Petrov, I.V. Zarutskiy, V.E. Kurochkin

Institute of Analytical Instrument Making, Russian Academy of Sciences, St. Petersburg, Russia
e-mail: alex.niispb@yandex.ru

Received December 16, 2021
Revised March 8, 2022
Accepted March 24, 2022

The success of genomic sequencing is impossible without the development of information technologies and mathematical methods for data processing to establish various features in the analyzed objects (nucleic acids) and trends in their changes. The volume of experimental data in the research of the genome has grown significantly, and new methods and algorithms are required for their processing. The primary stage of processing the data of devices for genomic parallel sequencing is the evaluation of the parameters of images obtained from video cameras in the form of electrical signals. The next stage of processing is the construction of a sequence of nucleotides according to algorithms that depend on the principle of operation of the device for sequencing nucleic acids. When performing this stage, algorithms for evaluating quality indicators for all individual readings (reads) are important. One of the ways to assess quality is to use algorithms based on the *k*-measure analysis methodology. The calculation of the number of occurrences of *k*-measures during the experiment on the parallel sequencing system makes it possible to assess the reliability of the analysis. In this article, algorithms for processing genetic analyzer data are considered.

Keywords: sequencing, nucleic acid, data processing, image processing.

DOI: 10.21883/TP.2022.07.54478.318-21

Introduction

DNA sequencing — is a method that allows to obtain information important for various human activity areas, including also synthetic biology, the essence of which is to determine sequential arrangement of nucleotides (A, T, G and C) in nucleic acid. Sequencing technology invention together with developed bioinformatics methods contributed a lot to organism genome analysis. Sequencer output file provides a sequence of four nucleotides. Maxam's and Gilbert's, and Sanger's developments constitute a milestone and opened opportunities for the development of more quick and efficient sequencing technology [1,2]. Modern sequencing, also referred to as next-generation sequencing technology (NGS), is characterized by very high capacity and much lower launch cost than those of previous technologies [3].

Operation of equipment for massive parallel sequencing (MPS) is based on the sequencing-by-synthesis technology — Solexa, composed of several stages:DNA fragmentation and adapter attachment, genomic library transmission through reaction cell channels coated with primers complementary to adapter ends, finishing of the second DNA strand by Bridge-PCR method (bridge amplification by polymerase chain reaction) followed by denaturation. After repeating two last actions, groups of identical molecules — clusters, are formed. Clusters are required to enhance optical signal.

The Institute for Analytical Instrumentation RAS (IAI RAS) is developing a hardware and software package (HSP) to decipher the nucleic acid sequence (NA) by „Nanophor SPS“massive parallel sequencing method. Data processing algorithms for the data obtained when HSP is running play an essential role in solution of genome sequencing tasks.

The purpose of this research is to analyze the capabilities of algorithms for processing of images generated in the genetic analysis process and for reliability assessment of the obtained nucleotide sequence when genomic sequencing tasks are solved. For reliability assessment of the obtained genetic data, *k*-mers incidence distribution analysis method is used. This method was used to process data obtained by the Russian-made „Nanophor SPS“sequencer.

1. Main fluorescent signal image processing operations

„Nanophor SPS“ parallel sequencing system uses four video cameras by the number of nucleotide types. Each of the video cameras is set to recording of one of the nucleotide types: A, C, G or T. Fluorescent signal is excited by two lasers in the pre-defined visible light range. The recorded light is transmitted through various light filters corresponding to the fluorescence wavelengths of each of the four dyes with which nucleotides are specifically labelled. Thus, each of the video cameras records the images of DNA molecule clusters on the end of which nucleotide with appropriate „letter“are arranged.

Due to structural features of the instrument, there are several technical difficulties in the achievement of full geometrical image coincidence of the same field of view of fluorescent objects. Image coordinate parameters of the

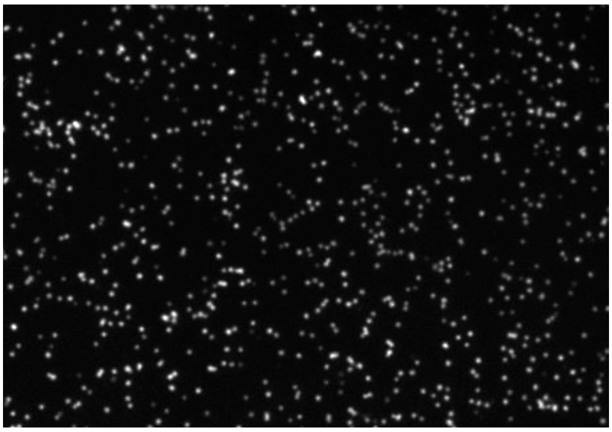

Figure 1. Initial fluorescent signal image for channel A (adenine).

same fluorescent object in different cameras may be shifted by several pixels. This research involved the development of algorithms and software for shift correction by mathematical methods.

The following operations are performed for fluorescent signal image processing in the parallel sequencing system:

1. Detection of background surface using morphological algorithms.

2. Image filtration using convolution with Mexicanhat.

3. Fluorescent object cluster detection algorithms.

4. Detection threshold identification algorithms.

5. Determining average fluorescent intensities of clusters.

6. Determining noise dispersion in image signals.

7. Correction of image shifts caused by structural features of the instrument by mathematical methods using cross-correlation functions.

8. Image focusing using lens movements.

Cluster density is an important parameter influencing successful sequencing result. Too low amount of clusters reduces the amount of output data. Too high amount of data causes „agglomeration“ of objects which affects data quality and results in experiment failure in some cases.

Performance algorithms for operations 1–7 are described in [4,5].

2. Image reading from video cameras

Four black-white video cameras set to recording of one of the nucleotide types: A, C, G or T. Cameras make it possible to record images with 4096 brightness gradations. Images from cameras arrive to the computer in the form of rasters — binary word arrays. Each word contains a brightness code of the corresponding pixel. Figure 1 shows the fluorescent signal image fragment for channel A — adenine.

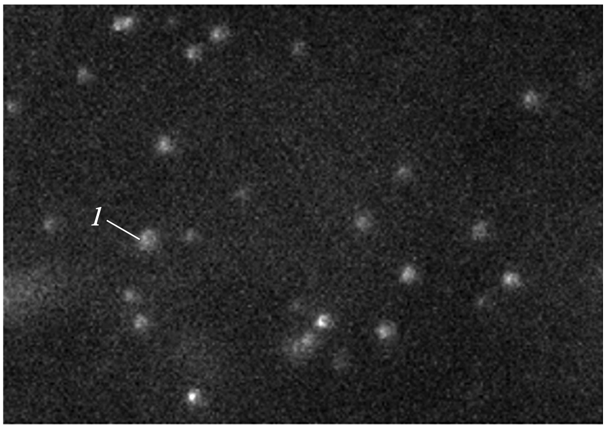

Figure 2. Field-of-view fragment image at the initial lens position with low focusing quality before focusing.

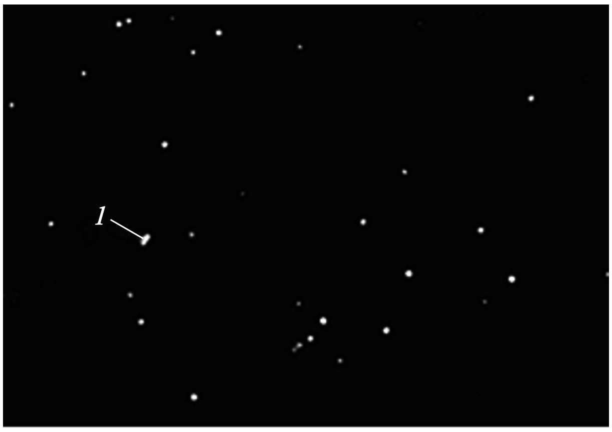

Figure 3. Field-of-view fragment image corresponding to the best focusing quality after focusing.

3. Blurred image focusing

In order to obtain focsed images of fluorescent objects, two methods are used: focusing by mathematical methods without lens position change and focusing using mechanical lens movement software.

Image focusing by mechanical lens movement involves determining of such lens position at which image focusing quality would be the highest and which will be described below.

Figure 2 and 3 show image fragments before and after focusing.

To check signal-to-noise ratio for images with low and high focusing quality, the signal amplitude was assessed for the object shown in Figure 2 and 3 with number *1*. Signal amplitude for this object on the image with low focusing quality was equal to 200 conventional units, and the signal amplitude for this object on the image with high focusing quality was equal to 400 conventional units. Noise level was determined on the basis of root mean square (rms) signal value in that area of the images where no objects were available. RMS values for images for images with low and high focusing quality were approximately equal to 10 conventional units. Thus, the signal-to-noise ratio for images with low and high focusing quality were equal to 20 and 40, respectively. For images shown in Figure 2 and 3, no maximum settings were used. For more clear visual assessment of the image with high focusing quality, Figure 3 is shown with high brightness.

Focusing procedure performance was studied as follows. There were sets of photographs of various reaction cell areas obtained with different lens positions, i. e. with various focusing degree. Each set contained photographs made in the same conditions (exposition and lighting). Then, for each of the photographs, focusing quality was assessed.

Focusing quality assessment methods have been being developed long before, a lot of them are available, but they are poorly studied, for example [6–10]. For operations described herein, 30 methods [6] were compared in relation to the images obtained using „Nanophor SPS". For choosing the best quality assessment method, functions were plotted where image number with different focusing was plotted on the x-axis and quality assessment by the appropriate method was plotted on the y-axis. The obtained functions were compared by three criteria:

1. The function shall have one extremum.

2. The function shall provide extremum for the image with the best focusing.

3. Robustness, i. e. gradient of function vs. image focusing degree.

Absolute focusing function values are not relevant, because only extremum position and gradient are relevant. For comparison, all function values were set to function maximum in extremum point, i. e. maxima were reduced to unity. To assess the image focusing quality, whole image is used, no individual part (line, column or field) is outlined.

The best results in terms of the criteria listed above were given by Vollath′s correlation [7] and dispersion methods [10]. The „Nanophor SPS" software uses dispersion for focusing quality assessment.

Figure 4 and 5 shows profiles of one of the fluorescent objects before and after focusing, respectively. After focusing, signal amplitude increased and profile width reduced.

For additional image focus correction and separation of partially overlapping signals of fluorescent objects, peaking filtration based on the inverse point-spread function convolution problem was used.

Blurred (defocused) image is mathematically obtained by means of point-spread function (PSF) convolution[11–14] with initial image. To ensure high quality reconstruction of the initial image, it is important to have the information about the true point spread function parameters. For blurred image reconstruction, inverse problem solution methods described in V.S. Sizikova et al. were used [12–14]. Mathematical reconstruction methods for blurred images make it possible to avoid mechanical lens focusing and reduce analysis time.

4. Determining image shifts in various channels and in various fields of sight using cross-correlation functions

For the purpose of this research, algorithms and software were developed for correction of image shifts caused by structural features of the instrument. These algorithms are based on calculation of cross-correlation functions between images obtained by different video cameras or between images from one video camera, but obtained in different experiment cycles (scans). Coordinates of maximum correlation function correspond to coordinates of test image shift.

Fluorescent signals under the dye action are generated for each of the nucleotides in the specific visible light wavelength (band) range. However, a phenomenon exists such as band overlapping between various nucleotide signals. This

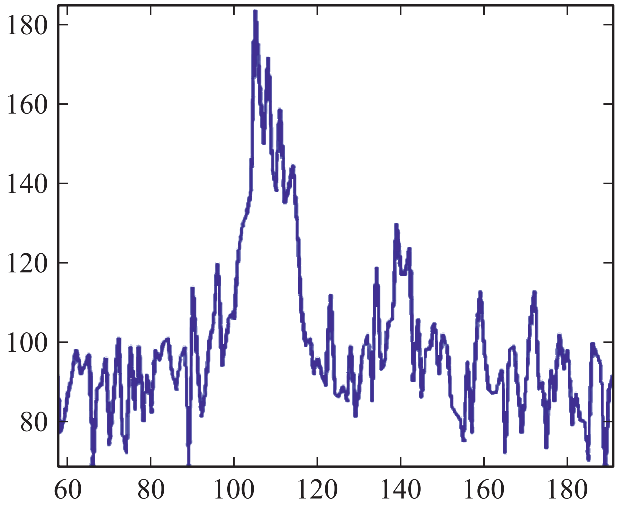

Figure 4. Profile of one of the defocused fluorescent objects shown in Figure 2. x-axis — pixel number on the image, y-axis — fluorescence intensity in conventional units.

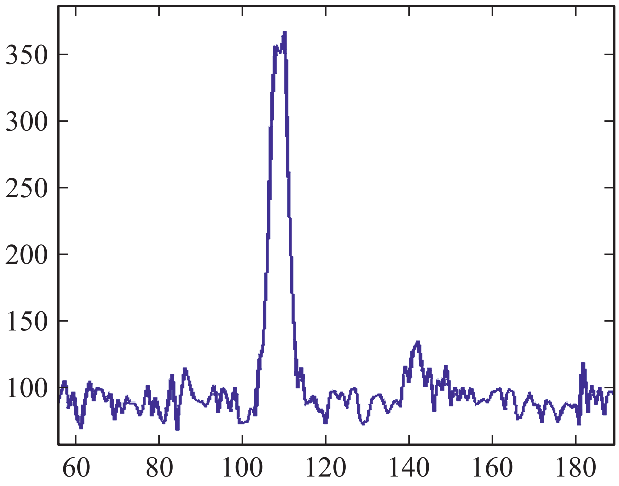

Figure 5. Fluorescent object profile shown in Figure 3 after focusing.

phenomenon means that the clusters which, for example, were„lit“ in channel A will be also „lit“ in channel C. This property is used for assessment of geometrical image shifts of different channels from each other. Image shift between different channels is defined using the cross-correlation function.

The cross-correlation function between two images was calculated using equation (1) [15]:

$$Y(u,v) = \frac{\sum\limits_{x,y}[f(x,y)-\bar{f}_{u,v}][g(x-y,y-u)-\bar{g}_{u,v}]}{\left\{\sum\limits_{x,y}[f(x,y)-\bar{f}_{u,v}]^2\sum\limits_{x,y}[g(x-y,y-u)-\bar{g}_{u,v}]^2\right\}^{0.5}}, \quad (1)$$

where $f(x,y)$ is a two-dimensional function of the first image; $g(x,y)$ is a two-dimensional function of the second image; x, y are image pixel coordinates; u, v are cross-correlation function coordinates, $\bar{f}_{u,v}$, $\bar{g}_{u,v}$ are average values of functions f and g, respectively.

Define the origin of coordinates of the two-dimensional image in maximum point of cross-correlation function of two similar images $u = 0$, $v = 0$. Now, consider the maximum cross-correlation image function coordinates. Maximum cross-correlation function of various image channels, for example „a“ and „c“, has coordinates corresponding to the sought-for shift. For example, the maximum cross-correlation function coordinates may be equal to $x = -2$, $y = 8$.

Similar calculations are carried out for shift coordinates for other channels and for single channel image shift, but recorded in different experiment cycles.

In addition to cross-correlation function calculations using equation (1), the image processing software use a faster calculation algorithms based on the forward and inverse fast two-fold Fourier transformation. Two-fold Fourier transformations of functions $f(x,y)$ and $g(x,y)$ are calculated. Then, Fourier-transforms of these functions are multiplied and a cross-correlation function is derived according to the Plancherel theorem [16].

5. Background correction

Initial image background means an image every pixel of which contains the signal intensity data without desired signal. The background correction algorithms uses a digital filter based on the first image convolution with the second derivative of the bivariate Gaussian function. This algorithms is simpler that the background correction algorithm using the digital low-pass filter based on the morphological erosion and dilatation operations described in [4]. Algorithm based on the convolution with the second derivative of the Gaussian function (Mexicanhat) reduces the background value virtually to zero, and the algorithm based on morphological operations reduces it be approx.5−7 times. Gaussian function width which is the basis for calculation of the second derivative is equal to 0.7 of the average width of the test image objects.

Figure 6 shows fluorescent signal profiles before background correction operation, after background correction on the basis of morphological operations and using convolution algorithm. It can be seen that the peak profile is at the background level consisting of about 120−140 conventional units. After background correction operation using the convolution algorithm, the background level is almost equal to zero. Negative signal values generated after convolution operation do not influence the assessment accuracy of fluorescent object center coordinates and their intensities, but may deteriorate the secondary data processing results at the nucleotide sequencing stage. Such values are removed during the secondary processing similarly to Swift software [17].

As shown in [4], the convolution algorithm allows noise filtering and „peaking“ in order to increase the coordinate assessment accuracy of the detected clusters. Noise reduction and „peaking“ by means of convolution algorithms are shown in Figure 7.

Convolution method with second derivative of Gaussian peaks is used in many image processing software of the type similar to „Nanophor SPS“.

6. Detection and assessment of local fluorescent object coordinates

After background subtraction operations, „peaking“ image filtration is used on the basis of convolution with the second derivative of the bivariate Gaussian function. Detection of objects is the operation of identification of image areas assigned to the sought-for objects. This is an essentially threshold operation, therefore it is important

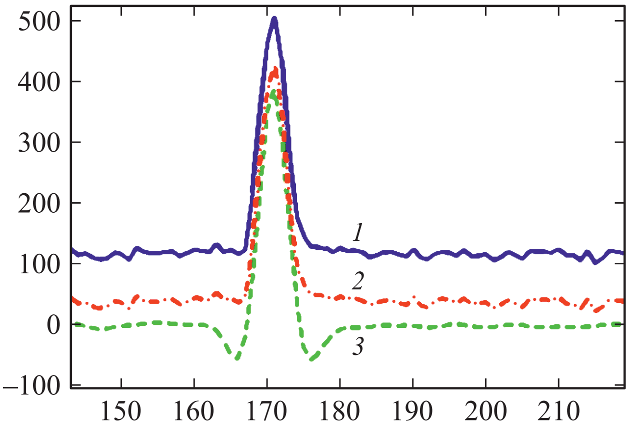

Figure 6. fluorescent object profile before and after background correction operations: *1* solid line — initial signal profile, *2* dotted line — signal profile after background correction using algorithms with morphological operations, *3* dashed line — signal profile after background correction using convolution algorithm. x-axis — pixel number on the image, y-axis — fluorescence intensity in conventional units.

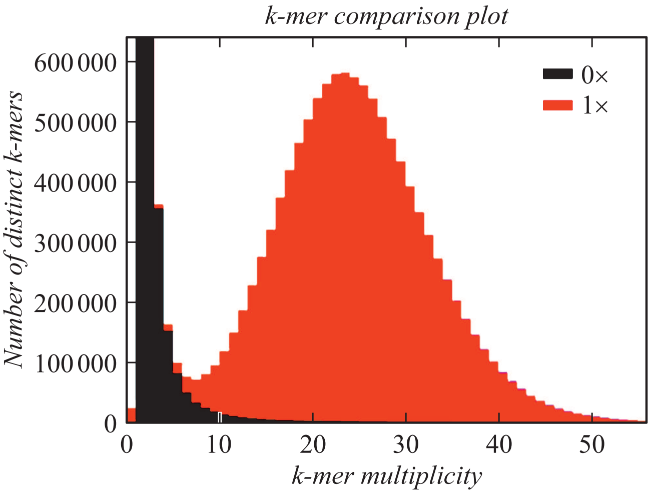

Figure 7. Near-perfect example of use of K-mer obtained using Nanophor SPS.

to determine the threshold which would allow to reliably separate the „signal" (object) from interference. To find the threshold, a signal intensity distribution bar chart method is used.

After the „peaking" filtration operation, not all agglomerated objects may be separated. To find and further separate the agglomerated objects, repeated peaking filtration operation is used, but with a more narrow core, and then a set of morphological processing operations is performed.

As a result of threshold processing, a bit map is generated which is a binary image. Pixels belonging to the objects in this image are equal to zero, and the background is equal to unity. The object coordinates are defined on the basis of maximum intensity in the detected object. The area is defined by the number of pixels exceeding the half of maximum intensity.

To define the boundaries, coordinates and areas, algorithms described in [4,5] were used.

7. „False" cluster rejection

One of the main operations for template creation [4] is superposition of binary images of separate fluorescent signal channels, for example, nucleotide channels „a" and „c" using binary operation OR. Such superposition takes place after image shift correction described above. After shift correction operation, errors of cluster coordinates on x and y are possible in come image points due to the fact that the obtained shift values are different for all image points. These errors do not exceed 1 or 2-x pixels, however, cause „false" clusters in the template and the nucleotide sequencing error is increased. To reduce such error, a routine is introduced in the fluorescent signal image processing software which performs „false" cluster rejection. If for any image point with coordinates x and y in a square area with coordinates in four angles, respectively, $(x-2, y-2)$, $(x+2, y-2)$, $(x-2, y+2)$, $(x+2, y-2)$, more than one cluster is detected, then all but one cluster are „false". „False" cluster rejection software leaves only one cluster in the template among all clusters detected in the square specified above. This cluster gives the maximum fluorescent signal intensity among all intensities detected in the specified cluster square.

Cluster coordinates and nucleotide sequences

Cluster number	H	v	nucleotide sequence
1	336	403	GACTGGTATTCCGCACCAGGTCTGGCCA
2	216	262	TTGTCCATTAGGCCCCACAAGGGCGGG
3	128	385	CCGTCGTCGTTACGGCCCCCGATAGTCG
4	5	16	GCTATGGATGCCCGGTCGCCGGCCCCA
5	266	398	AAGAGGGGTCTGGTCTCTTCACGGGCCT

8. Final result of fluorescent signal image processing software

The final result of the fluorescent signal image processing software is the table (see the fragment below) containing the cluster object image coordinates of the test nucleic acid. Each object of the detected cluster contains two coordinates: y — horizontal coordinate and x — vertical coordinate on the test images. Each cluster shall have a corresponding nucleotide sequence in the form of Latin letters detected for it: A, C, G, T (see the Table below).

9. Algorithms of reliability assessment of the obtained nucleotide sequence according to the assessment of *k*-mer incidence rate

K-mer is just a sequence of k-symbols in a string (or DNA sequence nucleotides in sequencing problem). For example, to get all k-mers from the sequence, get the first k-symbols, then shift to one symbol for beginning of the next k-mer, etc.

Sequence decomposition into its k-mers for analysis, makes it possible to analyze this set of fragments with a fixed size rather than the whole sequence and this may be a more effective approach. Operations with multiple k-mers are performed quicker and easier. A simple example is: in order to check whether sequence S originates from organism A or from organism B, assuming that genomes A and B are known and are different enough, we can check whether S contains more k-mers existing in A or in B.

Almost any genome contains repeated segments, however, beginning from a definite k, k-mers identify it uniquely. If we count the number of k-mer emergencies for a sufficiently high k (limited by read length on top), it turns that most of them are present as a single instance in the genome. For example, if the genome length order is comparable with human genome, the probability of meeting an unexpected substring with length 14 at least once is 0.975893 [18]. For $k = 20$ the same probability is 0.000909. For shorter genomes, e.g. bacteria and fungi, lower k may be chosen in order to achieve a similar low probability of multiple string incidence.

Analysis of k-mer incidence rate distribution allows to find assembly errors in already formed contigs [19,20]. Contig is a set of overlapping DNA-fragment sequences obtained from a single biological source (organism, tissue, cells).

10. *k*-mer handling procedures

In [18], genome assembly quality assessment method is proposed which allows to determine the correspondence between unique k-mers in the assembled genome and k-mers in reads. Read is a nucleotide sequence in a single cluster.

Procedure is as follows: 1. Construct a k-mer incidence bar chart for reads obtained during sequencing of the studied genome.

2. Choose some peak vicinity of unique k-mers in k-mer incidence bar chart in reads.

3. Construct k-mer incidence bar chart for each of the obtained assemblies.

4. Calculation of Q mer as a part of various k-mers taken rom the peak vicinity in k-mer incidence bar chart in reads among unique k-mers for assembled genome). Each k-mer from a set of unique k-mers of reads is checked whether it is included in a set of unique k-mers of the assembled genome.

In [21], an error correction method is proposed which is optimized for handling reads containing both substitution errors and insertion/ removal errors. Since errors occur with low probability, probability that the same k-mer is read several times with the same set of errors is very low. This means that those k-mers which occur in the set of reads a few times, are erroneous („bad“), the other are real genome substrings („good“).

K-mer spectrum — is a graphical presentation of a set of data showing how many short words with a fixed length (k-mers) appear for a certain number of times. Incidence rate is plotted in x-axis, and number of k-mers is plotted on y-axis.

We represent how many elements of each rate in the nucleotide (read) sequence spectrum are: 0 — not included in the reference genome (Phix174 in our case), included 1 once, 2 twice, etc.

Figure 7 shows a near-perfect example obtained using Nanophor SPS. No assembly errors are present (black peak outside of red one). One unique content is present in a reference set once.

Conclusion

The described algorithms and software developed on their basis record and handle fluorescent object signal images of „Nanophor SPS“parallel sequencing system. Among these algorithms and software, an algorithms for automatic image focusing, automatic assessment of image shift in various sequencing channels and cycles, background correction, field-of-view cluster coordinate detection and assessment software, etc.

To achieve the best automatic focusing quality of images obtained by video cameras, image pixel amplitude dispersion calculation criterion was selected as the most simple in software implementation.

The use of coordinate calculation algorithm for maximum two-dimensional cross-correlation function of two images makes it possible to perform automatic image shift assessment for images obtained by video cameras in various sequencer channels.

Convolution algorithm with the second derivative of Gaussian function allows noise filtration, „peaking“ and background influence correction.

Nucleotide sequencing, quality assessment algorithms are important for all individual reads. One of the quality assessment ways is the use of algorithms based on k-mers.

Methods based on k-mers shall be preferably used for efficient creation of genome assemblies.

KAT (k-mer Analysis Toolkit) software can analyze sequencing data to determine accident error, standard error level and contamination. Information obtained during the analysis may be helpful to make a decision whether further tasks such as genome assembly shall be continued. Then KAT can recheck this genome assembly by determining the assembly completeness and accuracy without any external reference data.

Combination of these algorithms in a single complex makes it possible to solve a set of important practical and research tasks for nucleotide sequencing of the test genome of various objects. The obtained results are useful in the following applications: molecular biology, genetics, agriculture, medicine, environment protection, etc.

Funding

The research was performed under State Assignment of the Ministry of Science and Higher Education of the Russian Federation № 075-00280-21-00 topic № 0074-2019-0013

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] F. Sanger, S. Niclein, A.R. Coulson. Proc. Natl. Acad. Sci. USA, **74**, 5463−5467 (1977). DOI: 10.1073/pnas.74.12.5463

[2] A.M. Maxam, W. Gilbert. Proc. Natl. Acad. Sci. USA, **74** (2), 560–564 (1977) DOI: 10.1073/pnas.74.2.560

[3] D.V. Rebrikov, D.O. Korostin, E.S. Shubina, V.V. Ilinskiy. *NGS: vysokoproizvoditel'noe sekvenirovanie*, pod obshchey red. D.V. Rebrikova (BINOM, Laboratoriya znaniy, M., 2014), 232 p. (in Russian)

[4] V.V. Manoilov, A.G. Borodinov, I.V. Zarutsky, A.I. Petrov, V.E. Kurochkin. Zhurn. Trudy SPII RAN, **18** (4), 1010–1036 (2019).(in Russian) DOI: 10.15622sp.2019.18.4.1010-1036

[5] V.V. Manoilov, I.V. Zarutsky. *Obrabotka signalov fluorestsentsii massovogo parallel'nogo sekvenirovaniya nukleinovykh kislot.* Sv-vo o gos. registratsii programmy dlya EVM. № 2019663248.

[6] S. Pertuz, D. Puig, M.Á. García. Pattern Recognition, **46** (5), 1415–1432 (2012). DOI: 10.1016/j.patcog.2012.11.011

[7] A. Santos, C.O. de Solorzano, J.J. Vaquero, J.M. Pena, N. Mapica, F.D. Pozo. J. Microscopy, **188**, 264–272 (1997).

[8] Yu Sun, S. Duthaler, B.J. Nelson. Microscopy Res. Tech., **65**, 139–149 (2004).

[9] Chun-Hung Shen, H.H. Chen. Robust Focus Measure for Low-Contrast Images. 2006 Digest of Technical Papers Intern. Conf. Consumer Electron., 69–70 (2006). DOI: 10.1109/ICCE.2006.1598314

[10] E. Krotkov, J.-P. Martin. Range From Focus. Proceed. IEEE Intern. Conf. Robotics and Automation, 1093–1098 (1986). DOI: 10.1109/ROBOT.1986.1087510

[11] R. Vuds, R. Gonsales. *Tsifrovaya obrabotka izobrazheniy* (Tekhnosfera, M. 2012), 3-e izd., ispr. i dop., 1104 p. (in Russian)

[12] V.S. Sizikov. *Pryamye i obratnye zadachi v vosstanovleniya izobrazheniy, spekroskopii i tomografii c Matlab* (Lan', SPb., 2017), 412 p. (in Russian)

[13] V.S. Sizikov. J. Opt. Technol., **84** (2), 95–101 (2017).

[14] V.S. Sizikov, A.V. Stepanov, A.V. Mezhenin, R.A. Burlov, R.A. Éksemplyarov. J. Opt. Technol., **85** (4), 95–101 (2018).

[15] J.P. Lewis. *Fast Template Matching, Vision Interface*, 120–123 (1995).

[16] Zh. Maks. *Metody i tekhnika obrabotki signalov pri phizicheskikh izmereniyakh: v 2-kh tomakh*, per. s frants. (Mir, M., 1983), vol. 1, 312, p. (in Russian).

[17] N. Whiteford, T. Skelly, Ch. Curtis, M.E. Ritchie, A. Löhr, A.W. Zaranek, I. Abnizova, C. Brown. Bioinformatics, **25** (17), 2194–2199 (2009). DOI: 10.1093/bioinformatics/btp38

[18] K. V. Romanenkov. *Metod otsenki kachestva cborki genoma na ocnove chastotk-merov*, Preprinty IPM im. M.V. Keldysha, 2017, 11.

[19] T.J. Treangen, D.D. Sommer, F.E. Angly, S. Koren, M. Pop. Current Protocols in Bioinformatics, **11** (11.8), 1–18 (2011).

[20] A.V.. Aleksandrov, A.A. Shalyto. Nauchno-tekhnichecky vestnik informatsionnykh tekhnologiy, mekhaniki i optiki, (1), 108–114 (2016) (in Russian)

[21] G. Marcais, C. Kingsford. Bioinformatics, **27** (6), 764–770 (2011).