

# The method of generation barcode for DNA certification of plants and organisms

O.Yu. Kiryanova<sup>1</sup>, I.I. Kiryanov<sup>2</sup>, L.U. Akhmetzianova<sup>1,3</sup>, B.R. Kuluev<sup>4</sup>, A.V. Chemeris<sup>4</sup>

<sup>1</sup>Ufa State Petroleum Technological University, Kosmonavtov street 1, Ufa, Russia, 450062

<sup>2</sup>Corning Inc, Shatelena street 26a, Saint-Petersburg, Russia, 194021

<sup>3</sup>Institute of Petrochemistry and Catalisys, Ufa Federal Research Center RAS, Prospect Oktyabrya 141, Ufa, Russia, 450075

<sup>4</sup>Institute of Biochemistry and Genetics, Ufa Federal Research Center RAS, Prospect Oktyabrya 71, Ufa, Russia, 450054

**Abstract.** In the current paper a new DNA certification method for living organisms were presented. The proposed approach is based on unique barcode that identifies a particular organism. The studies were conducted using several types of crops and model plant (potato, wheat, arabidopsis). The web based application was developed on the base of the proposed technique.

## 1. Introduction

Polymerase chain reaction (PCR) is an experimental method of molecular biology that can significantly increase the concentration of small DNA fragments in a sample [1]. PCR is widely used in biological and medical practice to isolate new genes, diagnose diseases and other tasks. There is copies accumulation of a specific nucleotide sequence.

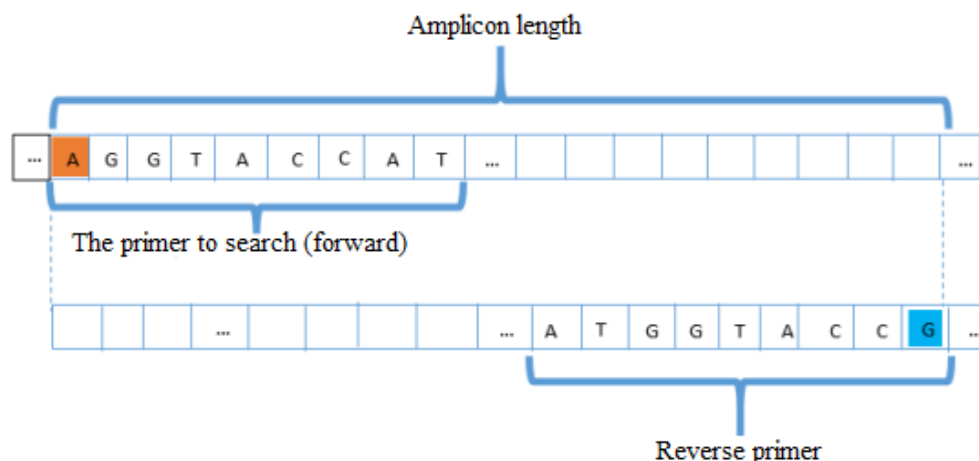
PCR was invented in 1983. Today it is the leading method in the field of physical and chemical biology.

Primers (short fragments consisting of 18-30 nucleotides) are important component that affects experiment success [2]. Primers in PCR must satisfy the main requirements: specificity of amplification process and efficiency. The size of human genome is about three billion nucleotides. Manual search of a random sequence with length of 18-30 nucleotides in such genome is very time consuming task. Suppose that any nucleotide is one character. The approximate number of characters in the novel «War and Peace» (including spaces and punctuation) is about 2 709 700 characters. Therefore, it is needed to determine whether fragment of text about 30 characters long is found in 1107 “War and Peace” novels. In case of standard PCR the common case is to define the presence of only one primer in a genome sequence. However, for multiplex PCR the search is performed for several primers (usually up to 10 and more). Multiplex PCR – PCR which involves more than one pair of oligonucleotide primers at the same time leads to the coamplification of several DNA matrix [3].

In order to solve this problem a web based application was developed. This program allows to determine the annealing positions of primers to the DNA matrix chain resulting from the length of amplicons. Since the probability of obtaining identical results for different genomes is negligible, the data obtained could be represented as unique barcode which, in its turn, could be a digital DNA passport [4].

## 2. Problem description

In order to determine the amplicon size it is necessary to know the position of the direct and reverse primers in a nucleotide sequence. After that the distance between these primers is determined. That distance is called size of the amplicon and must have 50...500 nucleotides length inclusive. This range is the most optimal for most cases of sequencing gel electrophoresis. A search example is shown on figure 1.



**Figure 1.** An example of searching forward and reverse primers in a fragment of nucleotide sequence.

Following the above-mentioned logic the proposed software collects information on all available occurrences of primers and amplicon lengths [5]. The result is represented in table form (table 1).

**Table 1.** Positions of annealing forward and reverse primers in the potato genome (data is placed in ascending order of amplicon size).

GGATCTTT position	AAAGATCC position	Amplicon size
39883835	39884052	217
55375264	553775548	284
29569657	29569969	312
38393029	38393375	346
49519668	49520023	355
41540764	41541163	399
8231987	8232448	461

Information about genome is presented as a single file or collection of files with text data according FASTA standard. This is the most common format for digital storage of nucleotide sequences. In other words, the text-based data storage format. Nucleotide sequences are stored as strings of characters “A”, “G”, “C”, “T” and “N”. Each letter means the corresponding nucleobase: adenine, guanine, cytosine, and thymine respectively. “N” means unknown nucleotide. FASTA-format allows easy data manipulations with sequences using text editors and programming languages such as Python, Ruby, Perl, etc.

The search places of annealing primers in a nucleotide sequence is implemented using the Boyer-Moore algorithm [6]. This algorithm is considered as the fastest among general-purpose algorithms designed to find a substring in a string. The advantage of this algorithm is that it costs a certain amount of preliminary calculations on the pattern (but not over the line where search is conducted). The pattern is not compared with the source text in all positions, most of them are skipped as obviously unsuccessful. General evaluation of the computational complexity of the Boyer-Moore algorithm –  $O(m+n)$ , where  $m$  – the length of the search pattern,  $n$  – length of the search string [7].

On the base of data about the length of amplicons a barcode is generated. The barcode is represented as a set of lines which determine the presence of amplicon length in the range from 51 to 500 nucleotides. We assumed that this range includes 450 conditional DNA cells, which may contain DNA (and this will be DNA<sup>+</sup>-cell) or no DNA (DNA<sup>-</sup>-cell). The presence of one or more DNA fragments with the same size in a specific DNA<sup>+</sup>-cell is not important since it is a qualitative rather than quantitative analysis. Thus, the information about each sample can be presented from alternating zeros and ones in the selected range of lengths taken in the amplicon analysis. For example, consider the range from 101 to 110 nucleotides, where the finding of DNA fragments has the following form: ...101<sup>+</sup>, 102<sup>+</sup>, 103<sup>-</sup>, 104<sup>-</sup>, 105<sup>-</sup>, 106<sup>-</sup>, 107<sup>+</sup>, 108<sup>-</sup>, 109<sup>+</sup>, 110<sup>-</sup>.... The numbers denote the size of DNA fragment in nucleotides, (+) – presence of a DNA fragment, (-) – absence of a DNA fragment. In binary format the entry for this section will be as follows: ...0100001000.

Visually such data can be conveniently represented as genetic barcodes in a linear or two-dimensional display. For example, for the data in table 1, the corresponding barcode is shown on figure 2.



**Figure 2.** Barcode example.

The main advantage of the proposed approach is the easier comparison of two independent genetic characteristics. It is possible to accurately measure the amplicon length after its separation in capillary gel electrophoresis under denaturing conditions.

The obtained data about the primer(s), the analyzed genome, and the set of selected amplicons are unique. It is completely eliminating the accidental barcode coincidence of different samples of strains, races, varieties, breeds, or individuals. Since the amplicons can have a huge number of variants (combinations) of the distribution of these DNA fragments on DNA<sup>+</sup> cells.

The total number of occurrences combinations in such DNA cells could be calculated as the number of combinations from  $m$  to  $n$  using the following formula (1):

$$C_m^n = \frac{m!}{n!(m-n)!} \quad (1)$$

where  $C$  is the total number of probabilistic occurrences combinations in DNA cells,  $m$  the number of all DNA- cells analyzed in the selected range and  $n$  the number of all DNA<sup>+</sup>-cells.

According to the probability theory, the largest number of combinations occurs when half of the cells are occupied with DNA fragments (225 of 450). In this case the number of combinations will exceed  $10^{100}$ . This number is more than enough for unambiguous DNA certification of any organism. The probability of a random match of two DNA samples with the number of different-sized amplicons equal to 5 will be about one case per  $10^{12}$ . Thus, the proposed approach is an efficient method for DNA certification of cultivars, lines, breeds, strains, and races.

### 3. ABCDNA\_GS (Amplified Bar-Coded DNA Genome/Specimen)

We have developed the web application with database for storing information about the amplicons and barcode generation.

Input data is: domain, Kingdom, genome, primer(s), type of DNA amplification.

The output data is: found amplicons sizes and the corresponding barcode.

As a result, found amplicons sizes allow to estimate the outcome of any particular PCR experiment.

In other words, obtained data allows to plan the PCR experiment for any genome.

User interface example is shown in the figure 3.

In addition to computer analysis, you can also populate experimental PCR data. Along the data population, DNA barcode is also could be generated from *in vivo* found amplicons. Moreover, the

generated information is a kind of digital passport for varieties, breeds, races, strains of various organisms.


## ABCDNA\_GS

Domain:  Kingdom:

Genome:

Primer(s):  Amplicons size:

DNA amplification method:

Barcode: 

**Figure 3.** The program interface, an example of the input data.

### 4. Conclusions

We have proposed a new approach for cataloging/certifying diverse plants. This unambiguous certification and identification is carried out by assigning unique genetic barcodes to plant varieties based on the detected DNA polymorphism. In addition, this method is applicable for all living organisms. Currently, many approaches are used for DNA certification of plant varieties but none of them provides unambiguous digital data. Thus, the developed approach technology of DNA certification (cataloging)/identification of living organisms is unique. In addition, the web application was developed that allows to detect the presence of specific primers in the DNA (genomes), determine the size of amplicons that are formed as a result of PCR, and create the corresponding unique barcode. The entire genomes of different organisms including wheat, potato and arabidopsis available from resource EnsembleGenomes <http://ensemblgenomes.org>. Thus, without conducting a full-scale experiment it is possible to test several primers as well as get an idea of the full-scale experiment success. The uniqueness of the proposed technique is that allows to systematize data for different primers and DNA sequences without taking into account their natural affiliation.

### 5. Acknowledgments

The work was supported by an RFBR grant 17-44-020120.

### 6. References

- [1] Glik, B. Molecular biotechnology. Principles and application / B. Glik, G. Pasternak – M.: Mir, 2002. – P. 589.
- [2] Garafutdinov, R.R. Variety of PCR primers and principles of their selection / A.K. Baimiev, G.V. Maleev, Ya.I. Alekseev, V.V. Zubov, D.A. Chemeris, O.Yu. Kiryanova, I.M. Gubaydullin,

- R.T. Matniyazov, A.R. Sakhabutdinova, Yu.M. Nikonorov, B.R. Kuluev, A.K. Baymiev, A.V. Chemeris // *Biomics*. – 2019. – Vol. 11(1). – P. 23-70.
- [3] Chamberlain, J.S. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification / J.S. Chamberlain, R.A. Gibbs, J.E. Ranier, P.N. Nguyen, C.T. Caskey // *Nucleic Acids Research*. – 1988. – Vol. 16(23). – P. 11141-11156.
- [4] What is FASTA format? – URL: <https://zhanglab.ccmb.med.umich.edu/FASTA/> (14.11.2019).
- [5] Kiryanova, O.Yu. Program for searching primers for polymerase chain reaction / O.Yu. Kiryanova, L.U. Akhmetzianova, B.R. Kuluev, I.M. Gubaydullin // *Materials of the XIII Russian scientific Internet conference « Integration of science and higher education in the field of bio-and organic chemistry and biotechnology» Ufa, 2019.* – P. 153-154.
- [6] Cormen, T.H. Algorithms: construction and analysis / T.H. Cormen, Ch.E. Leiserson, R.L. Rivest, K. Stein – M.: Williams, 2005. – P. 801.
- [7] Gusfield, D. Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology / D. Gusfield, I.V. Romanovskij – Spb.: Nevskij Dialekt, 2003. – P. 654.