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

Olga Kiryanova
Ufa State Petroleum Technological University
Ufa, Russia
olga.kiryanova27@gmail.com

Ilya Kiryanov
Corning, Inc.
Saint Petersburg, Russia
ilya.lsc@gmail.com

Liana Akhmetzianova
Institute of Petrochemistry and Catalisys;
Ufa Federal Research Center, RAS
Ufa, Russia
www.lianab@mail.ru

Bulat Kuluev
Institute of Biochemistry and Genetics;
Ufa Federal Research Center, RAS
Ufa, Russia
kuluev@bk.ru

Alexey Chemeris
Institute of Biochemistry and Genetics;
Ufa Federal Research Center, RAS
Ufa, Russia
chemeris@anrb.ru

Abstract—In the current paper a new DNA certification method for living organisms was presented. The suggested approach is based on unique barcode that identifies a particular organism. The studies were conducted using several species of crops and model plants (*Solanum tuberosum, Triticum aestivum, Arabidopsis thaliana*). The web based application was developed on the base of the proposed technique.

Keywords—polymerase chain reaction, primer design, DNA certification, barcode, web application

I. Introduction

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

PCR was invented in the midle of the 1980s. Nowadays it is the leading method in the field of physical and chemical biology.

Primers (short DNA fragments consisting of 10-30 nucleotides) are important components that affect on success of experiments [2]. Primers in PCR must satisfy the main requirements: specificity of amplification process and its efficiency. A pair of primers are usually used in PCR. However, in some cases a single primer may be sufficient since it is involved in forward and reverse primers simultaneously [3]. Such approach with single primer is used for DNA polymorphism elucidation. For multiplex PCR, several primers can be used simultaneously, usually up to 12. More than one pair of oligonucleotide primers at the same time leads to the coamplification of DNA matrices with results in multiple PCR products [4]. In this case primers could be annealed in pairs in all possible combinations. An example of primer annealing in the multiplex PCR is shown on figure 1.

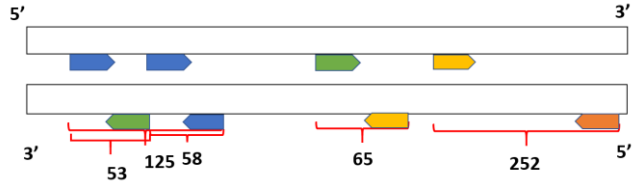

Fig. 1. An example of primer annealing in the multiplex PCR.

Different primers are shown in different colors. Red brackets denote the amplicons sizes.

It is possible to make predictions of amplicons sizes on the base of known complete nucleotide sequence of the analyzed organism. This is a complicated task which could not be done manually. For example, a genome with 1 billion pairs of nucleotides has about 10^3 annealing sites for decamer primers. To solve this problem a web based application was developed. The proposed software allows to determine the annealing positions of primers in the DNA chain indicating the length of amplicons. Since the probability of obtaining identical results for different genomes is negligible, the obtained data could be represented as unique barcode which, in its turn, represents a digital DNA passport [5].

II. Problem Description

The global efforts in creation and promotion of new varieties of agricultural crops requires the modernization of the selection process. Currently existing solutions for DNA certification of plants do not allow to obtain digitized data. The proposed barcode system is based on the polymorphism of specific genes (most often the cytochrome oxidase gene). Therefore, the detected degree of polymorphism is quite low and allows us to detect only the relationship of individual groups of organisms, as well as their location on the evolutionary tree [6-8]. Some recently dispersed species may not be distinguishable based on analysis of several genes. Modern instrumental methods for unambiguous genetic identification of biological material do not allow to determine the difference between plant varieties. The development of a well-reproducible and relatively inexpensive method of DNA certification of varieties and their DNA identification is an urgent task. Improved or new solutions for the abovementioned problem could ensure significant economic growth in the agricultural sector of economy.

For unambiguous certification and identification we proposed a new approach: to assign unique genetic barcodes to plant varieties based on the detected DNA polymorphism using PCR. It does not require prior knowledge about genome of any plant species.

There are more than 20 methods for detecting DNA polymorphism in plants. However, none of them could provide true digital data and does not have proper reproducibility [9-12]. The experimental basis of the DNA certification method is a modified PCR based on the RAPD -

Random Amplified Polymorphic DNA amplification method. It is preferable to perform computer analysis before the laboratory experiments conduction. Such computer modeling could assist to determine the places of possible annealing sites and sizes of reaction products (amplicons).

In order to determine the amplicon size *in silico* it is necessary to know positions of direct and reverse primers in a nucleotide sequence. After that the distance between these primers could be determined. That distance is called the amplicon size and must have from 51 to 500 nucleotides length inclusive. This range is optimal for most cases of gel electrophoresis and sequencing [13]. A search example is shown on figure 2.

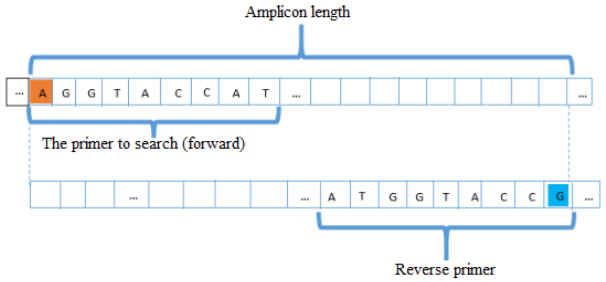

Fig. 2. An example of searching forward and reverse primers in a fragment of nucleotide sequence.

Following the above-mentioned logic the proposed software collects information on all available occurrences of primers and amplicons lengths [14]. An example of result is shown in table 1.

TABLE I. POSITIONS OF ANNEALING FORWARD AND REVERSE PRIMERS IN THE *SOLANUM TUBEROSUM* GENOME (DATA IS PLACED IN ASCENDING ORDER OF AMPLICON SIZE). EXAMPLE OF OUTPUT DATA .

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. Nucleotide sequences are stored as strings of characters “A”, “G”, “C”, “T” and sometimes “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. That is why FASTA files are widely used for primers positions search. According to the FASTA file format specification, above mentioned task could be reduced to the well-known approach: substring search in a string.

There are several well-known algorithms for substring search in a string: linear search, Knuth-Morris-Pratt algorithm and the Boyer-Moore algorithm [15]. The Boyer-Moore algorithm is considered as the fastest among general-purpose classical algorithms designed to find a substring in a string. The main advantage of the Boyer-Moore algorithm is that the shift is calculated based on the pattern (but not over the line where search is conducted). The pattern comparison with a fragment of the string occurs from right to left. In addition, the search 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 linear algorithm – $O(n \cdot m)$, where m is the length of the search pattern, n the length of the search string. General evaluation of the computational complexity of the Boyer-Moore and Knuth-Morris-Pratt algorithms is $O(m+n)$ [16].

A comparative analysis of algorithms efficiency was performed using genome with 10^6 nucleotides. It was shown that the Boyer-Moore algorithm is more suitable for primers search [17]. Thus, the Boyer-Moore algorithm was used to implement the substring (primer) search.

It should be mentioned, that one presented search technique was implemented using Python language with JIT Numba-compiler [18].

On the base of data about the length of amplicons a barcode could be 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 imaginary DNA cells, which may contain DNA (and this will be DNA^+-cell) or no DNA (DNA^--cell). The presence of one or more DNA fragments with the same size in a specific DNA^+-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^-, 102^+, 103^-, 104^-, 105^-, 106^-, 107^+, 108^-, 109^-, 110^- …. 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 could 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 3.

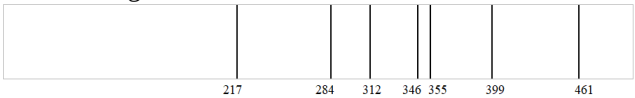

Fig. 3. Barcode example for *Solanum tuberosum* PCR reaction with forward primer GGATCTTT and reverse primer AAAGATCC.

The main advantage of the proposed approach is easy 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$^+$ 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$^+$-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 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 five is about one case per 10^{12}. Thus, the proposed approach is an efficient method for DNA certification of cultivars, lines, breeds, and strains.

III. 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 (Archaea, Prokaryotes, Eukaryotes), Kingdom (Animals, Plants, Fungus) – only for Eukaryotes, genome, primer(s), type of DNA amplification (RAPD, ISSR, AFLP). The entire genomes of different organisms including from resource EnsembleGenomes http://ensemblgenomes.org as FASTA files.

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. In addition, compare experimentally obtained amplicons with those found as a result of the program.

User interface example is shown on the figure 4.

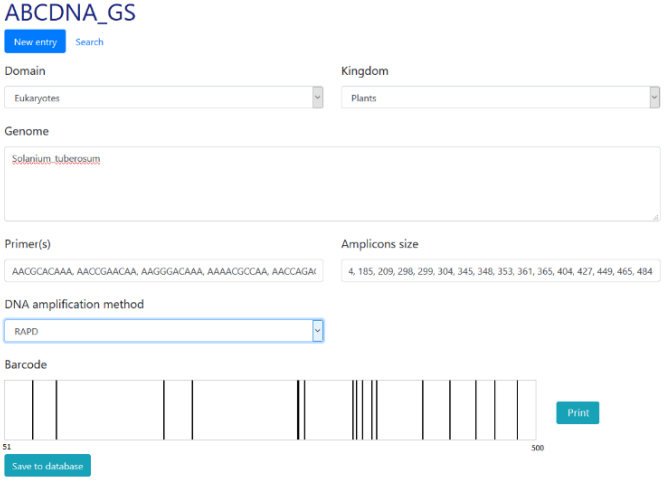

Fig. 4. The program interface, an example of the input data.

In addition to computer analysis it is possible to compare wet lab experiments (*in vitro* found amplicons) and predicted PCR outcome by comparing two barcodes.

Thus, the generated information is a kind of digital passport for varieties, breeds, strains of various organisms [19].

IV. Conclusions

We proposed a new approach for cataloging/certifying diverse groups of plants and other organisms. These unambiguous certification and identification were 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 besides human. Other methods are used for DNA identification of an individual approaches, the most promising for data barcode is considered to be single-nucleotide DNA polymorphism. Currently, many approaches are used for DNA certification of plant varieties but none of them provides unambiguous digital data. Thus, the suggested approach for 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. In the future, it is planned to translate data into QR code and use machine learning methods to classify barcodes and compare related varieties. [20].

Web based application allows to catalog wet laboratory experiments and *in silico* analysis. The entire genomes of different organisms including *Solanum tuberosum*, *Triticum aestivum*, *Arabidopsis thaliana* 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. Due to the uniqueness of the proposed approach it is possible systematize data for different primers and DNA sequences without taking into account their natural affiliation. It was shown that barcoding could enhance the genome comparison by excluding the human factor [21], allows to get digital data about a certain genome, and leads to the intuitive and clear comparison among other digitized genomes.

Acknowledgment

This research was supported by the Russian Foundation for Basic Research (project № 17-44-020120).

References

[1] B. Glik and G. Pasternak, “Molecular biotechnology. Principles and application,” Moscow: Mir, 2002, 589 p.

[2] R.R. Garafutdinov, An.Kh. Baymiev, 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, Al.Kh. Baymiev and A.V. Chemeris, “Variety of PCR primers and principles of their selection,” Biomics, vol. 11, no. 1, pp. 23-70, 2019.

[3] B.R. Kuluev, An.Kh. Baymiev, G.A. Gerashchenkov, D.A. Chemeris, V.V. Zubov, A.R. Kuluev, Al.Kh. Baymiev and A.V. Chemeris, “Random priming PCR strategies for identification of multilocus DNA polymorphism in eukaryotes,” Russian Journal of Genetics, vol. 54, no. 5, pp. 499-513, 2018.

[4] J.S. Chamberlain, R.A. Gibbs, J.E. Ranier, P.N. Nguyen and C.T. Caskey, “Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification,” Nucleic Acids Research, vol. 16, no. 23, pp. 11141-11156, 1988.

[5] "What is FASTA format?" [Online]. URL: https://zhanglab.ccmb.med.umich.edu/FASTA/.

[6] J. Huanga, Q. Xub, Z.J. Suna, G.L. Tanga and Z.Y. Sua, "Identifying earthworms through DNA barcodes," Pedobiologia, no. 51, pp. 301-309, 2007.

[7] A. Cywinska, S.L. Ball and J.R. deWaard, "Biological identifications through DNA bar-codes," Proc. R. Soc. Lond. B Biol. Sci., vol. 270, pp. 313-321, 2003.

[8] P.D.N. Hebert, S. Ratnasingham and J.R. deWaard, "Barcoding animal life: cytochromecoxidase I diver-gences among closely related species," Proc. R. Soc. Lond. B Biol. Sci., vol. 270, pp. 596-599, 2003. DOI: 10.1016/j.pedobi.2007.05.003.

[9] H. Nybom, K. Weising and B. Rotter, "DNA fingerprinting in botany: past, present, future," Investigative genetics, vol. 5, no. 1, 2014.

[10] K.N. Babu, M.K. Rajesh, K. Samsudeen, D. Minoo, E.J. Suraby, K. Anupama and P. Ritto, "Randomly amplified polymorphic DNA (RAPD) and derived techniques," Methods Mol Biol., vol. 1115, pp. 191-209, 2014. DOI: 10.1007/978-1-62703-767-9_10.

[11] N. Jones, H. Ougham, H. Thomas and I. Pasakinskiene, "Markers and mapping revisited: finding your gene," New Phytol., vol. 183, no. 4, pp. 935-966, 2009. DOI: 10.1111/j.1469-8137.2009.02933.x.

[12] P. Poczai, I. Varga, M. Laos, A. Cseh, N. Bell, J. P. Valkonen and J. Hyvönen, "Advances in plant gene-targeted and functional markers: a review," Plant Methods, vol. 9, no. 1, 2013. DOI: 10.1186/1746-4811-9-6.

[13] O.Yu. Kiryanova and A.V. Chemeris, "Modeling the search for primers in the DNA chain," Materials of the V International conference on information technology and nanotechnology ITNT, Samara, Russia, pp. 774-778, 2019.

[14] O.Yu. Kiryanova, L.U. Akhmetzianova, B.R. Kuluev and I.M. Gubaydullin, "Program for searching primers for polymerase chain reaction," 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, Russia, pp. 153-154, 2019.

[15] O.Yu. Kiryanova, L.U. Akhmetzianova and I.M. Gubaydullin, "Search algorithms in the analysis of nucleotide sequences for unambiguous identification of genomes," Bulletin of Bashkir University, vol. 25, no. 2, pp. 285-289. DOI: 10.33184/bulletin-bsu-2020.2.10.

[16] D. Gusfield, "Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology," Cambridge University Press, 2003, 654 p.

[17] T.H. Cormen, Ch.E. Leiserson, R.L. Rivest and K. Stein, "Algorithms: construction and analysis," M: Williams, 2005, 801 p.

[18] O.Yu. Kiryanova, I.I. Kiryanov, L.U. Akhmetzianova, B.R. Kuluev, A.V. Chemeris and I.M. Gubaydullin, "Parallel implementation of search algorithm for RNA guide design," Materials of International conference Parallel Computational Technologies (PCT), pp. 52-58, 2020.

[19] O.Yu. Kiryanova, I.I. Kiryanov, B.R. Kuluev, A.V. Chemeris, R.R. Garafutdinov and I.M. Gubaydullin, "ABCDNA_GS (Amplifaied Bar-Coded DNA Genome/Specimen)" [Online]. URL: https://www.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2020610703&TypeFile=html.

[20] V.V. Arlazarov, K. Bulatov, T. Chernov and V.L. Arlazarov, "MIDV-500: a dataset for identity document analysis and recognition on mobile devices in video stream," Computer Optics, vol. 43, no. 5, pp. 818-824, 2019. DOI: 10.18287/2412-6179-2019-43-5-818-824.

[21] B. Jiang, Y. Zhao, H. Yi, Y. Huo, H. Wu, J. Ren, J. Ge, J. Zhao and F. Wang, "PIDS: A User-Friendly Plant DNA Fingerprint Database Management System," Genes, vol. 11, no. 4, pp. 373, 2020.