ISSN 0003-6838, Applied Biochemistry and Microbiology, 2021, Vol. 57, No. 2, pp. 271–279. © Pleiades Publishing, Inc., 2021.
Russian Text © The Author(s), 2021, published in Prikladnaya Biokhimiya i Mikrobiologiya, 2021, Vol. 57, No. 2, pp. 196–205.

Use of Mathematical Methods for the Biosafety Assessment of Agricultural Crops

E. V. Korotkov[a, *], I. V. Yakovleva[a], and A. M. Kamionskaya[a]

[a] *Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Sciences, Moscow, 119071 Russia*

**e-mail: bioinf@yandex.ru*

Received June 17, 2020; revised October 30, 2020; accepted November 2, 2020

Abstract—In Russia and around the world, there are important questions regarding the potential threats to national and biological safety created by genetic technologies and the need to improve or introduce new, justified, and adequate measures for their control, regulation, and prevention. The article shows that a significant volume of the global market is occupied by five major transgenic crops, and producers are ready to switch to crops with an edited genome that has been approved in the United States, Argentina, and other countries. We propose a qualitatively new approach to the risk assessment of edited plants, "Safe Design," and we have also developed an extremely important, fundamentally new approach to the development of methods that combine next-generation sequencing (NGS) and Bioinformatics for the assessment of the crop import biosafety. The proposed mathematical approach provides a detailed analysis of the possible insertions of DNA fragments into the genome of edited crops and a clarification of their biological significance. The developed method can be used in the rapid screening of plants for the presence of potentially dangerous genes, viral sequences, and nonspecific promoter sequences.

Keywords: transgenic crops, gene-edited plants, safe design, biosafety, alignment, dynamic programming, full genomes, insertions, mutations

DOI: 10.1134/S000368382102006X

INTRODUCTION

The long-term strategy of Russia in genetic technologies requires fundamental research for analysis of the potential risks of new products in terms national and biological safety, as well as the need to improve or introduce new, justified, and adequate measures to control and regulate the identified risks. Obviously, the scale of biogenic threats accompanying new technologies does not have interstate borders [1]. Modern genomic and postgenomic technologies introduce previously nonexistent biotechnological products—new bioagents, including those of anthropogenic origin—into the biotechnological "landscape." Economic estimates of the "benefits–risks" ratio for the use of postgenomic biotechnologies show the legitimacy of their intensive implementation, since it promises a dramatic increase in the efficiency of the agroindustrial complex, which became necessary in the context of the economic crisis caused by the pandemic. However, as the volume and range of biotechnological (genetically engineered) products for various purposes increases, public concerns, both abroad and in Russia [2], are focused on problems that have not yet been resolved by science, such as

—the infeasibility of absolute biosafety of innovative technologies;

– threats of an unintentional or unauthorized release of biotechnological products into the environment (transgenic plants and animals, recombinant microorganisms);

– horizontal or vertical transfer of the transgene from biotechnological crops to unmodified analogs;

– uncontrolled leakage of genetic constructs into the environment during genetic engineering experiments or the production of recombinant products and other biothreats [3].

The use of nature-like genetic technologies for accelerated plant breeding, e.g., CRISPR/Cas technology, allows the introduction of mutations into the primary DNA sequence of the plant genome with previously unattainable accuracy and efficiency and opens up broad prospects for epigenetic changes. At the same time, this technological innovation raises questions about the inapplicability of specific current biosafety regulations for plants with an edited genome.

The concept of biological safety in the Russian legislative field is defined by several high-level documents. For example, the Decree of the President of the Russian Federation N 683 (2015 No. https://rg.ru/2015/12/31/nac-bezopasnost-site-dok.html) considers biosafety to be part of the national security norm (Article 3). The definition of biosafety is given in the

Russian GOST (GOST R 22.0.04-95 http://www.consultant.ru/cons/cgi/online.cgi?req=doc&base=EXP&n=267373#042176372250623007): “Biological safety is the state of protection of people, farm animals and plants, and the natural environment from the dangers caused by the source of the biological and social emergency.” More narrowly, in terms of genetically modified organisms (GMOs), GMO biosafety implies the absence of actual or predicted undesirable GMO effects (in comparison with the original unmodified organism) on the environment and on human and animal health.

The Russian legal environment ensuring the biosafety of the use of genetically modified plants is fragmented and very contradictory. Thus, Federal Law No. 358-FZ (2016) prohibits the cultivation and breeding of plants and animals with genetic code modified with genetic engineering methods or containing genetically engineered material that cannot be introduced as a result of natural processes. According to the article 7 of the Federal Law-149 “Seed production” (1997 http://www.consultant.ru/document/cons_doc_LAW_17121/), it is prohibited to import the seeds of such plants into the territory of the Russian Federation or to use them for sowing (planting). However, this prohibition on production in Russia does not apply to the import of genetically engineered food crops for use as food for the population and animal feed. The registration process successfully functions for food use for the population, and the registration of the same genetically modified crops as animal feed should be regulated by the Rules approved by the Government Decree No. 839, which are outdated and have not been enforced since 2013. At the moment, the import of GMOs of plant origin for use as feed is still regulated by separate decisions (PP No. 520, 2020 http://www.garant.ru/hotlaw/federal/1362141/#ixzz6PHkFMbJE), which is justified by the real need of the agro-industrial complex, e.g., GM soybean as a source of protein.

METHODOLOGICAL BACKGROUND AND PROBLEMS OF LEGAL REGULATION OF THE TURNOVER OF TRANSGENIC/GENE-EDITED AGRICULTURAL PLANTS

Consideration of the methodological background used to identify specific DNA sequences in plant tissues, plant raw materials, and products of their processing shows that it is based on the different variants of polymerase chain reaction: real-time PCR [4], matrix PCR (Order of the Ministry of Agriculture of Russia, 2017), and on a biological microchip [5].

Data from the International Service for the Acquisition of Agrobiotechnology Applications (ISAAA www.isaaa.org/kc/cropbiotechupdate/article/default.asp?ID=18166) demonstrates that five major GM (transgenic) crops occupied 99% of the world areas of GM agricultural crops in 2018: 38 GM soybean lines, 95.9 million ha; 137 GM maize lines, 58.9 million ha; 63 GM cotton lines, 24.9 million ha; 37 GM rapeseed lines, 10.1 million ha; and 5 GM alfalfa lines, 1.3 million ha. Since only 25 lines of plants of GMO origin were registered in Russia in 1999–2018, most GM lines remain outside the scope of registration, and, therefore, there are no systems for their identification. This leads to the following high risks: (a) loss of control over the unauthorized appearance of unidentified GMOs on the Russian market due to the backlog/lack of reference materials and technologies for the detection and identification of GMOs; (b) the penetration of the food and feed markets by unregistered imported GM products.

Analysis of the achievements in the production of gene-edited plants indicate that it is the risk from transboundary movement, not in classical transgenic plants but the production of new technologies, is increasing, e.g., gene-edited SDN-1 and SDN-2 crops that do not contain transgenes [6].

Thus, in June 2020, the Animal and Plant Health Inspectorate (APHIS) of the United States Department of Agriculture (USDA) approved an unregulated status for the HOLL soybean of the Calix company (Calyxt, United States, https://calyxt.com/calyxts-high-oleic-low-linolenic-soybean-deemed-non-regulated-by-usda/). It has a high oil content, the composition of which is characterized by a low content of linolenic acid. The HOLL soybean is the only commercialized product of the second generation. It has properties that are important for the consumer: higher stability and improved oil composition for the prevention of cardiovascular diseases. The HOLL soybean was obtained by the company Calyx with the TALENs® technology [7]. It contains only genetic material from the original organism (soybeans) with deletions in five target genes. This means that gene-edited soybean may appear on the U.S. market in two years, and then it will appear on the global market. The HOLL soybean produced by Calyx is one of the eight products that the company currently has under development, and it is expected that all of the products will receive USDA unregulated status (i.7, part 340. https://www.law.cornell.edu/cfr/text/7/part-340).

Genome-editing products, in many ways, respond to “big challenges” and in the future will provide many practical applications in a wide variety of areas, including traditional and nontraditional agricultural production, and will create fundamentally new products that do not have traditional analogs.

In Russia, the issue of excluding gene-edited, nontransgenic plants from the legislative environment related to GMOs is at the stage of scientific discussion and is being considered in the light of scientific and technical policy, economics, and societal interaction. An active discussion aimed at clarification of the similarities and differences between old and new risks inherent in already familiar and new biotechnological

Table 1. Ranking of regulatory restrictions on genomic editing in agriculture in selected countries

Country	Rating, score	Rating status
Brazil	10	Defined: no unique rules
Argentina	10	Defined: no unique rules
USA	10	Defined: no unique rules
Israel	8	Weakly regulated
Chili	5	Regulations under development
Paraguay	10	Defined: no unique rules
Japan	8	Weakly regulated
Canada	8	Weakly regulated
Australia	8	Weakly regulated
India	5	Regulations under development
Russia	5	Regulations under development
China	5	Regulations under development
UK	2	Mainly prohibited
EU	2	Mainly prohibited
Ukraine	1	Limited research, no defined rules

products and methods of their commercialization (release) has also begun abroad. The main argument for the "liberation" of the regulation of nontransgenic, gene-edited crops from GMO regulation is the fact that they are indistinguishable from plants created by traditional methods [8, 9]. It can be assumed that any risks associated with genome-editing products will be similar, equal, or lower than the risks associated with crops obtained by known breeding methods or already commercialized products [10, 11]. Table 1 shows the rating of countries based on the introduction of regulatory restrictions on the use of genome-editing products in agriculture in light of the current regulation in these countries [12]. It can be seen that the leading countries in the production of transgenic crops (Brazil, Argentina, United States) are also the world leaders in the genome editing of plants, especially in Argentina [13].

Table 2 shows some crops obtained with genome-editing technology that are undergoing testing procedures at some stage of registration.

Thus, it is clear that a wide range of edited crops will enter the global market in the coming years. In order to determine the prospects and potential of genome-editing and to enact a successful policy of responsible innovation, it is necessary to revise the methods for the identification, detection, and monitoring of new biotechnological products, in particular, a transition to the use of high-throughput sequencing and bioinformatics methods. The potential biohazards associated with the latest technologies must be countered by a qualitatively new approach, "Safe Design." This term refers to a process that can be defined as the implementation of a procedure for the identification and assessment of risks in the early stages of the design process, which will eliminate or minimize risks throughout the life of the created organism. The proposed approach includes the introduction of technical expertise during the earliest stage of in the planning/design/development of the created crop. In this regard, the practice of Argentina, which offers developers consultations in the "Regulator's Office" during the planning stage, is interesting [13]. Preventive risk assessment of newly edited crops, which is based on the accumulated scientific evidence, including long-term monitoring and field testing, is comparable to safety assessment in the aviation or nuclear industry and lays the groundwork for a "safety culture" in plant biotechnology.

It is possible to formulate a number of tasks that must be addressed in connection with the implementation of this proposal.

(1) The development of methods for the detection of genetically engineered manipulations in the genome of nontransgenic, biotechnological plant crops that could confirm the absence of nontarget mutations and the lack of expression of a new protein(s) possessing allergic and/or toxic properties due to a reading-frame shift.

(2) A demonstration of whether new biotechnological plants can pose any types of risks, in particular, different types of risks in comparison with already studied transgenic plants.

(3) The determination of potentially new areas of use for which new risks are understood or not well understood.

(4) The development of a model for the early risk assessment of gene-edited plants with the Safe Design principle. Increased security can be achieved with the inclusion of security parameters/functions in the

Table 2. Gene-edited crops

Crop	Development company, method	Status
Virus-resistant tomato https://www.nexgenplants.com/	Nexgen Plants (Australia)	Approved by USDA for field trials in 2017
Micro-tomato https://thecounter.org/international-space-station-gene-edited-tomato/	University of California, Riverside, United States CRISPR	For use on the International Space Station and in enclosed spaces and other confined spaces
Grapevine leafroll disease-resistant grapeshttps://geneticliteracyproject.org/2018/09/21/are-we-ready-for-genetically-modified-wine/	Rutgers University, USA, CRISPR	
Nonbrowning apples: varieties Arctic Golden, Granny Smith and Fuji https://www.arcticapples.com/how-did-we-make-nonbrowning-apple/	Okanagan Specialty Fruits, RNA interference.	Approved by USDA
Salt-tolerant rice https://www.forbes.com/sites/ariellasimke/2020/02/21/you-may-find-salt-tolerant-rice-growing-in-the-ocean-by-2021/?subId1=xid:fr1582662210931gjd#1c4569ce4133	Agrisea, editing	Floating ocean farms
High-fiber wheat https://calyxt.com/calyxt-harvests-high-fiber-wheat-field-trials/	Calyxt, editing	Approved by USDA for field trials in 2018
Camelina (Brassicaceae family) with improved oil composition (omega-3) https://www.nature.com/articles/nbt0118-6b.epdf?shared_access_token=SS4V7V5nwo6_VHeVnriWkNRgN0jAjWel9jnR3ZoTv0MwSccfXlkuLBszumLMvCj9t-ForwjJaKkVVBMsLKWESjOw0sSf21kBJtFPCTmLUrUKqgSmVJpXParouCNHw0Ww98VQyz5Rr-Fyg2BDc5u16A%3D%3D	Yield10 Bioscience CRISPR	Approved by USDA in 2017
Drought-tolerant and salt-tolerant soybean https://www.nature.com/articles/nbt0118-6b.epdf?shared_access_token=SS4V7V5nwo6_VHeVnriWkNRgN0jAjWel9jnR3ZoTv0MwSccfXlkuLBszumLMvCj9t-ForwjJaKkVVBMsLKWESjOw0sSf21kBJtFPCTmLUrUKqgSmVJpXParouCNHw0Ww98VQyz5Rr-Fyg2BDc5u16A%3D%3D	University of Minnesota, United States, CRISPR	Approved by USDA in 2017
High-yielding tomato with increased fruit and reduced branching and amount of leaves https://qz.com/989925/scientists-are-perfecting-salad-by-editing-mutated-tomato-genes/	Cold Spring Harbor Laboratory, editing	Developed in 2017
Improved quality alfalfa https://swseedco.com/press-release/calyxt-and-sws-gene-edited-alfalfa-plant-designated-as-non-regulated-by-usda/	Calyxt, TALEN	Nonregulated status by USDA in 2017
Wheat with improved resistance to powdery mildew https://www.nature.com/news/gene-editing-surges-as-us-rethinks-regulations-1.19724	Calyxt, TALEN	
Nonbrowning potato https://geneticliteracyproject.org/2016/10/27/calyxts-bruise-resistant-non-browning-gmo-potato-cleared-sale/	Calyxt, TALEN	Approved by USDA in 2016

Table 2. (Contd.)

Crop	Development company, method	Status
Corn with high starch content (waxy corn) https://www.washingtonpost.com/news/wonk/wp/2017/06/13/how-one-company-plans-to-change-your-mind-about-genetically-edited-food/	DuPont, CRISPR	Nonregulated status by USDA in 2016
Drought-resistant maize https://onlinelibrary.wiley.com/doi/full/10.1111/pbi.12603	DuPont, CRISPR	Developed in 2016
Champignon https://www.scientificamerican.com/article/gene-edited-crispr-mushroom-escapes-u-s-regulation/	Pennsylvania State University, United States, CRISPR	Nonregulated status by USDA in 2016

developmental standard for the security-certification protocol.

IDENTIFICATION OF VARIOUS ARTIFICIAL INSERTIONS OF DNA FRAGMENTS IN THE GENOME OF AGRICULTURAL PRODUCTS

The section will review the possibilities of bioinformatics for the identification of artificial rearrangements in the genome. Over the past 30 years, bioinformatics has developed a variety of computational methods to study the base sequences in DNA and RNA. The first task was a pairwise comparison of two sequences of bases of DNA and RNA or two amino-acid sequences. Today, the most complete solution is obtained via dynamic programming [14]. In this case, there are two sequences, and it is necessary to draw a conclusion about their similarity in the case of the substitution of nucleotides or amino acids, as well as their insertions or deletions in previously unknown places and of unknown length. Methods for global and local comparisons of sequences have been developed [15, 16]. With global alignment, two sequences are compared from beginning to end. Local alignment searches for fragments of two sequences with the best match. Heuristic programs of the Blast family and the Fasta program have also been developed [15, 17–19]. Although- these programs use heuristic algorithms, they can quite accurately find pairwise similarities between amino acid sequences.

This task was subsequently expanded for the comparison of various genomes. The so-called genomic browsers were created; they make it possible to compare not only relatively short sequences but complete genomes. The most popular of them include the University of California Santa Cruz (UCSC) Genome Browser and the Ensembl Genome Browser [20, 21]. Specialized programs for the comparison of sequences of complete genomes have also been developed [22, 23]. These tools allow the comparison of any plant or animal genome of the considered product with a genome that has already been sequenced. In addition, multiple comparisons of different genomes or specific regions can be made. Such a comparison of genomes allows the relatively simple detection of insertions or deletions of DNA fragments that are present in one genome but not in another genome. This also allows for the clear identification of point mutations (SNPs).

Simultaneously with the development of methods for the comparison of DNA sequences, databanks containing most of what has been sequenced in the world have been intensively developed. There are two main data banks: the EMBL databank [24] created by the European Molecular Biology Laboratory (formed by 20 member countries and the partner country Australia) and Genbank [25]. These databanks contain not only the sequences of various DNA fragments but also the sequences of various complete genomes. These genomes include the genomes of many bacteria, viruses, plants, and animals. Databanks describing in detail the genomes of individual organisms, e.g., Solanaceae species, and combining crops important for agriculture, such as tomato *Solanum lycopersicum,* potatoes *S. tuberosum*, pepper *Capsicum annuum* (https://www.solgenomics.net/), were created. Such databanks also include related wild species of these crops, which serve as donors of various agronomically valuable traits (e.g., resistance to abiotic and biotic stresses) during the selection of varieties.

The availability of such information makes it possible to find the reference genome for an agricultural plant (AP). The reference genome is the genome of a traditional AP that has already been previously sequenced and entered into the database; therefore, we can further consider two situations: the presence and absence of a reference genome.

The reference genome is present. In this case, if the DNA sequence of the AP genome is determined, it can then be compared with the reference genome with bioinformatic methods. As a result of this comparison, it is possible to identify a variety of insertions or deletions of DNA fragments that could be made in the AP genome, as well as SNPs in the genome. Thus, if reference genome and a sequenced AP genome are available, the problem of finding any undeclared insertions or deletions can now be solved. In this case, it is possible to provide full control over the AP biosafety at the genome level.

There is also the question of the types of insertions or deletions in the AP genome that possess the highest biological hazard. The most potentially dangerous types include some insertion of promoter sequences, the insertion of potentially dangerous genes, or the insertion of any viral sequences or their parts. The insertion of promoter sequences could change the expression profile of any genes in the AP genome, which can lead to changes in the processes of plant development. Thus, the AP can acquire new biological properties that were not present in a plant with a reference genome. If any genes are inserted, the potential hazard depends on the functional significance of the inserted gene. The most dangerous genes are those encoding a variety of toxins. DNA fragments belonging to various viruses of both humans and agricultural animals or plants can also be classified as dangerous insertions into the AP genome. Moreover, even plant viruses cannot be recognized as completely safe for humans and animals [26]. In this case, the use of such an AP can have serious consequences for the population or for the agriculture of the Russian Federation.

The reference genome is absent. The AP genome does not always have a reference genome. In this case, the algorithms for pairwise and multiple comparison of genomes will not identify artificially made rearrangements of the genome of a new agricultural product, since there is nothing to compare the AP genome with. Therefore, in addition to comparative analysis methods, it is necessary to develop mathematical methods and algorithms for the annotation of potentially dangerous genomic sequences. The annotation is the determination of the functional role of various AP sequences. In the AP genome, it is possible to search for sequences based on their functional significance. For such a search, it is necessary to create sets (a database) of biologically hazardous DNA sequences that should not be present in the AP genome. Such sequences include at least

(1) various viral sequences of both humans and APs or animals;

(2) gene sequences encoding bacterial toxins or toxins of any other origin;

(3) all promoter sequences that can be found around the genes listed in point 2.

The search for promoter sequences in AP genomes is important for two reasons. First, the discovery of promoter sequences indicates the location of possible genes, which can help in their identification. Second, promoter sequences in some cases make it possible to distinguish genes from pseudogenes and indicate the transcription start sites (TSSs). This means that the identification of promoter sequences can help to isolate those genes that can be transcribed in the AP. Such identification requires the sequence of the complete genome of the AP. Sequencing of the entire genome is a fairly financially expensive technology today, but the cost of sequencing is decreasing, and its widespread use will become possible in the coming years.

The question arises as to how to find the sequences listed in points 1–3 in the AP genome if the reference genome is missing. A standard approach involving the production of many DNA sequences (*M*) that perform the same biological function can be used to search for such sequences. This is the so-called training set. It is desirable to include in the training set such similarities that only occur in potentially biologically hazardous sequences. Thus, it is necessary analyze all existing databases of nucleotide sequences and create the *M* set for each potentially dangerous sequence from points 1–3.

Then, for each set *M,* a conditional Markov model (CMM) is created. It is used to search for DNA sequences that perform the same biological functions as sequences from the set *M* [27–29]. This CMM is "scanned" throughout the AP genome, and the sequences that are members of set *M* are identified. This means that a lot of the potentially dangerous sequences for set *M* were produced in the AP genome. Many genes and protein families have been annotated in this way. It is relatively easy to create set of different pathogens and viral sequences with the existing EMBL and Genbank databases. However, there are difficulties in this approach that cannot be solved with the existing mathematical methods.

The problem is that the use of CMM works fine as long as the sequences have not accumulated a sufficient number of insertions and substitutions of nucleotides or insertions or deletions of both single nucleotides and extended sequences. If the number of mutations per nucleotide between any pair of sequences in set *M* is more than 2.4 [30], then a statistically significant multiple sequence alignment cannot be constructed. This will make it impossible to search for sequences in the AP genome that have the same functional significance as sequences from set *M*, i.e., a sequence with a given biological function, e.g., a potential toxin, will exist in the AP genome, but it will not be possible to identify it with the existing approaches.

This leads to the need to develop new mathematical methods to identify various DNA sequences in the genomes of APs that do not have a reference genome. This means that complete biological safety of APS currently without a reference genome cannot be achieved.

A new method for the multiple alignment of highly divergent sequences (MAHDS) has been proposed for the full identification of possible insertions or deletions in the AP genome. On the site http://victoria.biengi.ac.ru/mahds/auth, this method can be used to construct multiple alignments for nucleotide sequences. Highly divergent sequences are sequences that have accumulated more than 2.5 random substitutions (x) per nucleotide relative to each other. MAHDS makes it possible to construct statistically significant alignments for x in the range from 2.4 to 4.4 (http://victoria.biengi.ac.ru/mahds/auth) [31]. It was shown [31] that previously developed algorithms can construct statistically significant multiple alignments up to $x < 2.4$.

MAHDS is currently used for the multiple alignment of promoter sequences from the *Arabidopsis thaliana* genome and the human genome [32]. This study showed that statistically significant multiple alignments for promoter sequences cannot be calculated with existing methods, since $x = 3.6$ for them [31]. Multiple alignments for 4220 promoter sequences from the rice genome were constructed with the MAHDS method. A method was also developed for the creation of promoter classes based on the performed multiple alignment. In total, it was possible to create five classes of promoter sequences with a class size of more than 100 promoters. The obtained classes of promoter sequences were used to search for other promoter sequences in the rice genome. A profile matrix with a size of (16.600) was created for each class [33, 34]. The search for potential promoter sequences was performed for each template with a global alignment. In total 145277 potential promoters were identified. Of these, 18563 were promoters of known genes, which accounted for about 46% of the annotated genes. An algorithm for the analysis of randomly mixed nucleotide sequences of the complete rice genome was applied to calculate the number of false positives. The number of false positives in this case was about 1×10^{-8} per nucleotide. If the inverted chromosome sequence from the rice genome was taken as a control and the developed algorithm was applied, the number of false positives was then 4×10^{-7}. In any case, this was significantly less than the values obtained with all of the methods used to search for promoter sequences in eukaryotic genomes.

The existing algorithms for the prediction of promoter sequences cannot plot statistically significant alignments for promoter sequences; therefore, other mathematical approaches are used. These include such algorithms as TSSW [35], PePPER [36], G4PromFinder [37], and many others. The best algorithms predict a false positive at a level of 10^{-3}–10^{-4} per nucleotide, while the rice genome contains ~4.3×10^8 DNA bases. As a result, it is impossible to isolate the real promoter among tens of thousands of false predictions. In fact, the search for promoter sequences with computer methods is currently only possible with the MAHDS method. However, MAHDS is only a computational method, and complete confirmation that the detected sequences are functional promoters is possible either with experimental methods or the study of the similarity of the revealed sequences with sequences of different transcriptomes. In the latter case, it is possible to detect TSS and to search for the similarity of DNA sequences located upstream of the TSS with potential promoter sequences revealed by the developed mathematical method.

At the same time, the correlation of the revealed potential promoter sequences with various dispersed repeats and transposons was studied. It was possible to show that ~87000 promoter sequences correlate with the various dispersed repeats and transposons found in a previous study [38]. A total of 20654 promoter sequences belong to the previously annotated rice promoters. In this case, the number of false positives was not higher than 160 sequences. The remaining 37390 potential promoter sequences may represent promoters of unknown genes (in particular, microRNA genes [39]), promoters associated with various mobile elements of the genome, or evolutionary traces of the dispersal of genes and their promoters. These promoter sequences are the most interesting from a biosafety point of view. The reason is that a small number of SNPs in the promoter sequence can transfer the gene located behind the promoter sequence from an inactive state to an active, and the active transcription of a previously silent gene can begin. In fact, this means that some of the biological properties of an AP can change from being completely safe to being potentially dangerous.

The MAHDS method is universal and can be used to construct multiple alignments for any nucleotide sequence. Such sequences can include the various viral sequences listed above or those for different toxic genes, animals, plants, and humans. It is only necessary to create the corresponding sets M of sequences and to construct multiple alignments for them with the MAHDS method. After that, the AP genome is “scanned” by each set M (it can be tens of thousands of such sets), and a reasoned conclusion on the presence of potentially dangerous sequences listed in paragraphs 1–3 is obtained. After that, rapid identification of the possible insertions of DNA fragments into the AP genome becomes possible, even for sequences that have accumulated a significant number of base substitutions and insertions or deletions.

CRISPR/Cas9-EDITED PLANTS

It should be noted that whole-genome sequencing and subsequent bioinformatic analysis of the obtained data, including genome assembly and annotation, is currently an expensive and laborious approach to the comparative assessment of AP genomes. In addition, recently developed methods of genome editing, such as CRISPR/Cas9, unlike other known methods (agrobacterial or biolistics transformation), allow the necessary modifications to be made without traces. Therefore, the development of new approaches to the determination of possible changes in a genome of artificial origin is becoming urgent. The basis for such an approach may be the creation of a databank of sequences associated with economically valuable traits of species and varieties of agricultural crops based on the vast amount of information available today [40, 41]. Further, with modern search methods [42, 43, https://crispr.cos.uni-heidelberg.de/], it is possible to identify in the selected sequences the sites that are most likely to be used for the design of the so-called guide RNA (determines the location CRISPR/Cas9-editing) and to collect them into a separate database. Thus, the search for insertions, deletions, stop codons (excluding gene expression or production of the correct protein), and nonsynonymous single nucleotide substitutions can be narrowed down to a comparative analysis of a set of short sequences in a number of genes associated with certain plant characteristics.

CONCLUSIONS

Whole-genome sequencing and bionformational methods open up unique, new opportunities for the assessment of crop biosafety. It has become possible to carry out a detailed analysis of possible insertions of DNA fragments in the AP genome and to determine their biological significance. The rapid screening of APs for the presence of potentially dangerous genes, viral sequences, and nonspecific promoter sequences is also possible. An almost complete identification of APs containing unwanted or biohazardous genes, oncogenes, and genes that produce toxins will also be possible. The application of these approaches in practice will significantly increase the AP biosafety.

FUNDING

This study was partial financially supported by the Russian Foundation for Basic Research, project no. 18-29-14067\18.

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that they have no conflict of interest. This article does not contain any studies involving animals or human participants performed by any of the authors.

REFERENCES

1. Korobko, I.V., Georgiev, P.G., Skryabin, K.G., and Kirpichnikov, M.P., *Acta Naturae*, 2016, vol. 8, no. 4, pp. 6–13.
2. Yakovleva, I.V., Zhuravleva, E.V., and Kamionskaya, A.M., *FEBS Open Bio*, 2019, vol. 9, no. S1, pp. 280–281.
3. Guidance on Risk Assessment of Living Modified Organisms, Convention on Biological Diversity, 2012. UNEP/CBD/BS/COP-MOP/6/13/Add.1 July 30, 2012.
4. *Metodicheskie ukazaniya MUK 4.2.1903-04. "Produkty pishchevye. Metod identifikatsii geneticheski modifitsirovannykh istochnikov (GMI) rastitel'nogo proiskhozhdeniya s primeneniem biologicheskogo mikrochipa", utverzhdennye Glavnym gosudarstvennym sanitarnym vrachom RF 6 marta 2004 g.* (Guidelines MUK 4.2.1903-04. "Food Products. Method for the Identification of Genetically Modified Sources (GMI) of Plant Origin using a Biological Microchip", Approved by the Chief State Sanitary Doctor of the Russian Federation, March 6, 2004), Moscow: Minzdrav Rossii, 2004. http://base.garant.ru/4181368/#friends.
5. *Metodicheskie ukazaniya MUK 4.2.1902-04. "Opredelenie geneticheski modifitsirovannykh istochnikov (GMI) rastitel'nogo proiskhozhdeniya metodom polimeraznoi tsepnoi reaktsii", utverzhdennye Glavnym gosudarstvennym sanitarnym vrachom RF 6 marta 2004 g.* (Guidelines MUK 4.2.1902-04. "Determination of Genetically Modified Sources (GMI) of Plant Origin by the Method of Polymerase Chain Reaction", Approved by the Chief State Sanitary Doctor of the Russian Federation, March 6, 2004), Moscow: Federal'nyi tsentr Gossanepidnadzora Minzdrava Rossii, 2004. http://base.garant.ru/4180376/.
6. Agapito-Tenfen, S.Z., Okoli, A.S., Bernstein, M.J., Wikmark, O.G., and Myhr, A.I., *Front. Plant Sci.*, 2018, vol. 9, pp. 1–18. Article 1874. https://doi.org/10.3389/fpls.2018.01874
7. Maher, M.F., Nasti, R.A., Vollbrecht, M., Starker, C.G., Matthew, D.C., and Voytas, D.F., *Nat. Biotechnol.*, 2020, vol. 38, pp. 84–89. doi.org/ https://doi.org/10.1038/s41587-019-0337-2
8. Jones, H.D., *Nat. Plants*, 2015, vol. 1. Article 14011. https://doi.org/10.1038/nplants.2014.11
9. Davison, J. and Ammann, K., *GM Crops Food*, 2017, vol. 8, no. 1, pp. 13–34. https://doi.org/10.1080/ 21645698.2017.1289305
10. Metje-Sprink, J., Mens, J., Dodrzejewski, D., and Sprink, T., *Front. Plant Sci.*, 2019, vol. 9, pp. 133–141. Article 1957. https://doi.org/10.3389/fpls.2018.01957
11. Globus, R. and Qimrom, U., *Cell Biochem. J.*, 2018, vol. 119, no. 2, pp. 1291–1298. https://doi.org/10.1002/jcb.26303
12. Yakovleva, I.V., Vinogradova, S.V., and Kamionskaya, A.M., *Russ. J. Genet.: Appl. Res.*, vol. 6, no. 6, pp. 646–656. https://doi.org/10.1134/S2079059716060095
13. Whelan, A.I. and Lema, M.A., *GM Crops Food*, 2015, vol. 6, no. 4, pp. 253–265. https://doi.org/10.1080/21645698.2015.1114698
14. Eddy, S.R., *Nat. Biotechnol.*, 2004, vol. 22, no. 7, pp. 909–910. https://doi.org/10.1038/nbt0704-909
15. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., *J. Mol. Biol.*, 1990, vol. 215, no. 3, pp. 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
16. Needleman, S.B. and Wunsch, C.D., *J. Mol. Biol.*, 1970, vol. 48, no. 3, pp. 443–453.
17. Pearson, W.R. and Lipman, D.J., *Proc. Natl. Acad. Sci. U. S. A.*, vol. 85, no. 8, pp. 2444–2448. https://doi.org/10.1073/pnas.85.8.2444
18. Mount, D.W., *CSH Orotocols*, 2007, vol. 2007. https://doi.org/10.1101/pdb.top16
19. States, D.J., Gish, W., and Altschul, S.F., *Methods*, 1991, vol. 3, no. 1, pp. 66–70. https://doi.org/10.1016/S1046-2023(05)80165-3

20. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., and Flicek, P., *Database*, 2016, vol. 2016. Article bav096.
https://doi.org/10.1093/database/bav096

21. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D., *Genome Res.*, 2002, vol. 12, no. 6, pp. 996–1006.
https://doi.org/10.1101/gr.229102

22. Derrien, T., Andre, C., Galibert, F., and Hitte, C., *Bioinformatics*, 2006, vol. 23, no. 4, pp. 498–499.
https://doi.org/10.1093/bioinformatics/btl618

23. Sinha, A.U. and Meller, J., *BMC Bioinformatics*, 2007, vol. 8. Article 82.
https://doi.org/10.1186/1471-2105-8-82

24. Hingamp, P., Broek, A.E., Stoesser, G., and Baker, W., *Mol. Biotechnol.*, 1999, vol. 12, no. 3, pp. 255–267.
https://doi.org/10.1385/MB:12:3:255

25. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W., *Nucleic Acids Res.*, 2013, vol. 41, no. D1, pp. D36–D42.
https://doi.org/10.1093/nar/gks1195

26. Nikitin, N.A., Trifonova, E.A., Karpova, O.V., and Atabekov, J.G., *Moscow Univ. Biol. Sci. Bull.*, 2016, vol. 71, no. 3, pp. 128–134.

27. Yoon, B.J., *Curr. Genomics*, 2009, vol. 10, no. 6, pp. 402–415.
https://doi.org/10.2174/138920209789177575

28. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer, E.L.L., *Nucleic Acids Res.*, 1999, vol. 27, no. 1, pp. 260–262.
https://doi.org/10.1093/nar/27.1.260

29. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., and Bateman, A., *Nucleic Acids Res.*, 2010, vol. 38, no. S1, pp. 211–222.
https://doi.org/10.1093/nar/gkp985

30. Korotkov, E.V. and Korotkova, M.A., *J. Phys.: Conf. Ser.*, 2019, vol. 1205. Article 012025.
https://doi.org/10.1088/1742-6596/1205/1/012025

31. Korotkov, E.V., Suvorova, Y.M., Kostenko, D., and Korotkova, M.A., *Genes*, 2021, Under consideration.

32. Korotkov, E.V., Kamionskaya, A.M., and Korotkova, M.A., *Biotekhnologiya*, 2020, vol. 36, no. 4, pp. 7–14.
https://doi.org/10.21519/0234-2758-2020-36-4-7-14

33. Pugacheva, V., Korotkov, A., and Korotkov, E., *Stat. Appl. Genet. Mol. Biol.*, 2016, vol. 26, no. 5, pp. 381–400.
https://doi.org/10.1515/sagmb-2015-0079

34. Suvorova, Y.M., Korotkova, M.A., Skryabin, K.G., and Korotkov, E.V., *DNA Res.*, 2019, vol. 26, no. 2, pp. 157–170.
https://doi.org/10.1093/dnares/dsy046

35. Solovyev, V.V., Shahmuradov, I.A., and Salamov, A.A., in *Methods in Molecular Biology*, Ladunga, I., Ed., N.J.: Humana Press, 2010, vol. 674, pp. 57–83.
https://doi.org/10.1007/978-1-60761-854-6_5

36. De Jong, A., Pietersma, H., Cordes, M., Kuipers, O.P., and Kok, J., *BMC Genomics*, 2012, vol. 13. Article 299.
https://doi.org/10.1186/1471-2164-13-299

37. Di Salvo, M., Pinatel, E., Tala, A., Fondi, M., Peano, C., and Alifano, P., *BMC Bioinformatics*, 2018, vol. 19. Article 36.
https://doi.org/10.1186/s12859-018-2049-x

38. Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., Jiang, N., Hirsch, C.N., and Hufford, M.B., *Genome Biol.*, 2019, vol. 20, no. 4. Article 275.
https://doi.org/10.1186/s13059-019-1905-y

39. Zhou, X., Ruan, J., Wang, G., and Zhang, W., *PLoS Comput. Biol.*, 2007, vol. 3, no. 3, pp. 0412–0423. e37.
https://doi.org/10.1371/journal.pcbi.0030037

40. Mohan, V. and Paran, I., in *The Capsicum Genome. Compendium of Plant Genomes*, Ramchiary, N. and Kole, C, Eds., New York: Springer Cham, 2019, pp. 105–109.

41. Vemireddy, L.R., Noor, S., Satyavathi, V.V., Srividhya, A., Kaliappan, A., Parimala, S.R.N., Bharathi, P.M., Deborah, D.A., Rao, K.V.S., Shobharani, N., Siddiq, E.A., and Nagaraju, J., *BMC Plant Biol.*, 2015, vol. 15. Article 207.
https://doi.org/10.1186/s12870-015-0575-5

42. Stemmer, M., Thumberger, T., del Sol., Keyer, M., Wittbrodt, J., and Mateo, J.L., *PLoS One*, 2015, vol. 10, no. 4. e0124633.
https://doi.org/10.1371/journal.pone.0124633

43. Labuhn, M., Adams, F.F., Ng, M., Knoess, S., Schambach, A., Charpentier, E.M., Schwarzer, A., Mateo, J.L., Klusmann, J.H., and Heckl, D., *Nucleic Acids Res.*, 2017, vol. 46, no. 3, pp. 1375–1385.
https://doi.org/10.1093/nar/gkx1268

Translated by V. Mittova