ПРИКЛАДНАЯ БИОХИМИЯ И МИКРОБИОЛОГИЯ, 2021, том 57, № 2, с. 196–205

УДК 608.3.: 316.422.42

ИСПОЛЬЗОВАНИЕ МАТЕМАТИЧЕСКИХ МЕТОДОВ С ЦЕЛЬЮ ОЦЕНКИ БЕЗОПАСНОСТИ СЕЛЬСКОХОЗЯЙСТВЕННЫХ КУЛЬТУР

© 2021 г. Е. В. Коротков[1], *, И. В. Яковлева[1], А. М. Камионская[1]

[1]Институт биоинженерии, Федеральный исследовательский центр "Фундаментальные основы биотехнологии" Российской академии наук, Москва, 119071 Россия

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

Поступила в редакцию 17.06.2020 г.
После доработки 30.10.2020 г.
Принята к публикации 02.11.2020 г.

В России и в мире остро стоит вопрос относительно создаваемых генетическими технологиями потенциальных угрозах национальной и биологической безопасности и необходимости совершенствования или введения новых, оправданных и адекватных мер контроля, регулирования для их предотвращения. В настоящее время значительный объем мирового рынка занимают 5 основных трансгенных культур, а их производители готовы к переходу на сельсхозяйственные культуры с редактированным геномом, одобренными в США, Аргентине и некоторых других странах. Предлагается качественно новый подход к оценке рисков редактированных растительных организмов – "Безопасное проектирование". Математический подход, крайне востребованный и принципиально новый подразумевает разработку методик, использующих сочетание полногеномного секвенирования и бионформационных методов для оценки биобезопасности сельскохозяйственных культур. Он делает реальностью детальный анализ возможных вставок фрагментов ДНК в геном редактированных сельскохозяйственных растений и выяснение их биологического значения. Разработанный метод может служить для быстрого скрининга растительных организмов на присутствие потенциально опасных генов, вирусных последовательностей и неспецифических промоторных последовательностей.

Ключевые слова: трансгенные растения, редактированные культуры, безопасное проектирование, биобезопасность, полные геномы, вставки, мутации, выравнивание, динамическое программирование

DOI: 10.31857/S0555109921020069

Долгосрочная стратегия России в области генетических технологий требует проведения фундаментальных исследований для анализа потенциальных рисков получаемой новой продукции с точки зрения национальной и биологической безопасности, а также необходимости совершенствования или введения новых, оправданных и адекватных мер контроля и регулирования выявленных рисков. Масштабы биогенных угроз, сопровождающих перспективные новейшие технологии, с очевидностью, не имеют межгосударственных границ [1]. Современные геномные и постгеномные технологии вносят в биотехнологический "ландшафт" ранее не существовавшие биотехнологические продукты – новые биоагенты, в том числе антропогенного происхождения. Экономические оценки соотношения "преимущества–риски" использования постгеномных биотехнологий показывают правомерность их интенсивного внедрения, так как это обещает резко повысить эффективность агропромышленного комплекса, что стало необходимо в условиях экономического кризиса, вызванного пандемией. Однако по мере нарастания объемов и спектра биотехнологических (генно-инженерных) продуктов разного назначения, опасения общества как за рубежом, так и в России [2] фокусируются на еще нерешенных наукой проблемах таких, как:

– недостижимость абсолютной биобезопасности инновационных технологий;

– угрозы непреднамеренного или несанкционированного выпуска биотехнологической продукции в окружающую среду (трансгенные растения и животные, рекомбинантные микроорганизмы);

– горизонтальный или вертикальный перенос трансгена от биотехнологических культур к немодифицированным аналогам;

– неконтролируемая утечка в окружающую среду генетических конструкций в ходе генно-инженерных экспериментов или производства рекомбинантных продуктов и другие биоугрозы [3].

Использование природоподобных генетических технологий для ускоренной селекции растений, на-

пример, технологии CRISPR/Cas, позволяет с недостижимой ранее точностью и эффективностью вносить мутации в первичную последовательность ДНК растительного генома, открывает широкие перспективы также и для эпигенетических изменений. Одновременно, эта технологическая новация ставит вопросы о неприменимости конкретных действующих нормативных положений, касающихся биобезопасности, к растениям с редактированным геномом.

Само понятие биологической безопасности в российском законодательном поле определено несколькими документами высокого уровня. Так, Указ Президента РФ N 683 (2015 г. https://rg.ru/2015/12/31/nac-bezopasnost-site-dok.html) рассматривает биобезопасность в составе нормы о национальной безопасности (статья 3). Определение биобезопасности дается в российском ГОСТ (ГОСТ Р 22.0.04-95 http://www.consultant.ru/cons/cgi/online.cgi?req=doc&base=EXP&n=267373#042176372250623007): **“Биологическая безопасность** – состояние защищенности людей, сельскохозяйственных животных и растений, окружающей природной среды от опасностей, вызванных или вызываемых источником биолого-социальной чрезвычайной ситуации”. Более узко, в части генно-инженерно-модифицированных организмов (ГМО), под **биобезопасностью ГМО понимается** отсутствие фактического или прогнозируемого нежелательного воздействия ГМО (в сравнении с исходным немодифицированным организмом) на окружающую среду, здоровье человека и животных.

Российское правовое поле в сфере обеспечения биобезопасности использования генетически модифицированных растений разрозненно и весьма противоречиво. Так, Федеральный закон № 358-ФЗ (2016 г.), запрещает выращивание и разведение растений и животных, генетическая программа которых изменена с использованием методов генной инженерии и которые содержат генно-инженерный материал, внесение которого не может являться результатом природных (естественных) процессов. А согласно статье 7 ФЗ-149 “О семеноводстве” (1997 г. http://www.consultant.ru/document/cons_doc_LAW_ 17121/), запрещается ввозить на территорию Российской Федерации и использовать для посева (посадки) семена таких растений. Однако этот запрет на производство на территории России не распространяется на импорт полученных с использованием генной инженерии продовольственных культур с целью использования в качестве пищевых продуктов для населения и кормов для животных. Регистрационный процесс успешно функционирует в сфере пищевого использования для населения, а регистрация тех же генетически модифицированных (**ГМ**) сельскохозяйственных культур в качестве кормов для животных должна регулироваться Правилами, утвержденными Постановлением правительства № 839, которые устарели, не вступив в действие с 2013 года. На настоящий момент ввоз ГМО растительного происхождения для использования в качестве кормов все еще регулируется отдельными решениями (ПП № 520, 2020 г. http://www.garant.ru/hotlaw/federal/1362141/#ixzz6PHkFMbJE), обоснованными реальной потребностью агропромышленного комплекса, например, в ГМ сое как источнике белка.

Методическая база и проблемы нормативно-правового регулирования оборота трансгенных/редактированных сельскохозяйственных растений. Рассмотрение методической базы, используемой для выявления специфических последовательностей ДНК в растительных тканях, растительном сырье и продуктах их переработки, показывает, что она основана на методе полимеразной цепной реакции в различных вариантах: в режиме реального времени [4], в матричном формате (Приказ Минсельхоза России, 2017), а также на основе биологического микрочипа [5].

Данные Международной службы по коммерциализации достижений агробиотехнологии (ISAAA www.isaaa.org/kc/cropbiotechupdate/article/default.asp?ID=18166) демонстрируют, что в 2018 г. 5 основных ГМ (трансгенных) культур занимали 99% мировых площадей генно-инженерно-модифицированных сельскохозяйственных культур: 38 линий ГМ сои – 95.9 млн. га, 137 линий ГМ кукурузы – 58.9 млн. га, 63 линии ГМ хлопчатника – 24.9 млн. га, 37 линий ГМ рапса – 10.1 млн. га, 5 линий ГМ люцерны – 1.3 млн. га. Учитывая, что в России за период 1999–2018 гг. зарегистрировано всего 25 линий ГМО растительного происхождения, большая часть ГМ линий остается за рамками регистрации, и, следовательно, отсутствуют системы их идентификации. Этим определяются следующие высокие риски: а) потери контроля за несанкционированным появлением на российском рынке не идентифицированных ГМО в связи отставанием/отсутствием референсных материалов и технологий обнаружения и идентификации ГМО; б) проникновения на продовольственный и кормовой рынки незарегистрированной ГМ продукции, поступающей по импорту.

Анализ достижений в области создания редактированных растений позволяет заключить, что возрастают риски трансграничного перемещения не классических трансгенных растений, а продукции новых технологий, например, редактированных культур SDN-1 и SDN-2, не содержащих трансгенов [6].

Так, в июне 2020 г. Инспекция здоровья животных и растений (APHIS) Государственного Департамента сельского хозяйства США (USDA) утвердила нерегулируемый статус для сои “HOLL” с высоким содержанием масла, состав

Таблица 1. Рейтинг нормативных ограничений геномного редактирования в сельском хозяйстве некоторых стран

Страна	Рейтинг, балл	Статус рейтинга
Бразилия	10	Определен: нет уникальных правил
Аргентина	10	Определен: нет уникальных правил
США	10	Определен: нет уникальных правил
Израиль	8	Слабо регулируемый
Чили	5	Правила в разработке
Парагвай	10	Определен: нет уникальных правил
Япония	8	Слабо регулируемый
Канада	8	Слабо регулируемый
Австралия	8	Слабо регулируемый
Индия	5	Правила в разработке
Россия	5	Правила в разработке
Китай	5	Правила в разработке
Великобритания	2	В основном запрещено
ЕС	2	В основном запрещено
Украина	1	Ограниченные исследования, нет четких правил

которой характеризуется низким содержанием линоленовой кислоты, компании Каликс (“Calyxt”, США https://calyxt.com/calyxts-high-oleic-low-linolenic-soybean-deemed-non-regulated-by-usda/). Соя “HOLL” – единственный коммерциализованный продукт, который относится к продуктам второго поколения, то есть обладает свойствами, важными для потребителя: более высокой стабильностью и улучшенным составом масла для профилактики сердечно-сосудистых заболеваний. Соя “HOLL” была получена с использованием технологии TALENs® в варианте компании “Каликс” [7], она содержит генетический материал только исходного организма (сои) с делециями в пяти целевых генах. Это означает, что соя с редактированным геномом может появиться на рынке США через два года, а затем – и на мировом рынке. Соя “HOLL” компании “Каликс” – один из 8 продуктов компании, находящихся на данный момент в стадии разработки, и, как предполагается, они все получат нерегулируемый статус USDA (i.7, part 340. https://www.law.cornell.edu/cfr/text/7/part-340).

Продукция геномного редактирования, во многом, отвечает на “большие вызовы” и в перспективе обеспечит множество практических приложений в самых разнообразных областях, включая традиционное и нетрадиционное агропроизводство, создаст принципиально новые продукты, не имеющие традиционных аналогов.

В России вопрос об исключении редактированных нетрансгенных растений из законодательного поля, относящегося к ГМО, находится в стадии научного обсуждения, и рассматривается также в свете научно-технической политики, экономики, взаимодействия с обществом. За рубежом также развернута активная дискуссия по выяснению сходства и различия старых и новых рисков, присущих уже привычным и новым биотехнологическим продуктам и способам их коммерциализации (выпуска). Основным аргументом за “освобождение” из-под ГМО регулирования не содержащих трансгенов редактированных культур, является тот факт, что они не отличимы от растений, созданных традиционными способами [8, 9]. Можно предположить, что и любые риски, связанные с продуктами геномного редактирования, будут аналогичными, равными или меньшими, чем риски, связанные культурами, полученными известными методами селекции, или уже коммерциализованными продуктами [10, 11]. В табл. 1 приведен рейтинг стран по введению нормативных ограничений использования продуктов геномного редактирования в сельском хозяйстве в свете действующего в этих странах регулирования [12]. Из этих результатов видно, что страны-лидеры по производству трансгенных культур (Бразилия, Аргентина, США) вырвались и в мировые лидеры геномного редактирования растений, особенно это относится к Аргентине [13].

В табл. 2 приведены некоторые сельскохозяйственные культуры, полученные технологией геномного редактирования, которые проходят процедуры испытаний и находятся на той или иной стадии регистрации.

Таким образом, ясно, что в ближайшие годы в оборот на мировой рынок поступит широкий ассортимент редактированных сельскохозяйственных культур. Чтобы раскрыть перспективы и потенциал геномного редактирования и проводить успешно политику ответственных инноваций, необходим пересмотр методов идентификации, детекции и мониторинга новой биотехнологической продукции, в частности, переход к использованию методов высокопроизводительного секвенирования и биоинформатики. Потенциальным биоопас-

Таблица 2. Редактированные сельскохозяйственные культуры

Культура	Компания-разработчик, метод	**Статус**
Устойчивый к вирусам томат https://www.nexgenplants.com/	Nexgen Plants (Австралия)	Одобрен USDA для начала полевых испытаний в 2017 г.
Микро-томат https://thecounter.org/international-space-station-gene-edited-tomato/	Университет Калифорнии, Риверсайд, США CRISPR	Для использования на Международной космической станции, а также в закрытых помещениях и других ограниченных пространствах
Виноград, устойчивый к вирусу скручивания листьев https://geneticliteracyproject.org/2018/09/21/are-we-ready-for-genetically-modified-wine/	Университет Рутгерс, США, CRISPR	
Яблоко, нечернеющее : Сорта Arctic Golden, Granny Smith и Fuji https://www.arcticapples.com/how-did-we-make-nonbrowning-apple/	Okanagan Specialty Fruits, РНК-интерференция	Одобрены USDA
Солеустойчивый рис https://www.forbes.com/sites/ariellasimke/2020/02/21/you-may-find-salt-tolerant-rice-growing-in-the-ocean-by-2021/?sub-Id1=xid:fr1582662210931gjd#1c4569ce4133	Agrisea, редактирование	Плавучие океанские фермы
Пшеница с высоким содержанием клетчатки https://calyxt.com/calyxt-harvests-high-fiber-wheat-field-trials/	Calyxt, редактирование	Одобрена USDA для полевых испытаний в 2018 г.
Камелина (растение из семейства горчичных) с улучшенным составом масла (омега-3) https://www.nature.com/articles/nbt0118-6b.epdf?shared_access_token=SS4V7V5nwo6_VHeVnriWkNRgN0jAjWel9jnR3ZoTv0MwSccfXlkuLBszumLMvCj9t-ForwjJaKkVVBMsLKWESjOw0sSf21kBJtFPCTmLUrUKqgSmVJpXParouCNHw0Ww98VQyz5Rr-Fyg2BDc5u16A%3D%3D	Yield10 Bioscience CRISPR	Одобрена USDA в 2017 г.
Засухоустойчивая и солеустойчивая соя https://www.nature.com/articles/nbt0118-6b.epdf?shared_access_token=SS4V7V5nwo6_VHeVnriWkNRgN0jAjWel9jnR3ZoTv0MwSccfXlkuLBszumLMvCj9t-ForwjJaKkVVBMsLKWESjOw0sSf21kBJtF-PCTm LUrUKqgSmVJpXParouCNHw0Ww98VQyz5Rr-Fyg2BDc5u16A%3D%3D	Университет Миннесоты, США, CRISPR	Одобрена USDA в 2017 г.
Высокоурожайный томат с увеличенным количеством плодов и уменьшенным ветвлением, и количеством листьев https://qz.com/989925/scientists-are-perfecting-salad-by-editing-mutated-tomato-genes/	Лаборатория Колд-Спринг-Харбор, редактирование	Разработан в 2017 г.
Люцерна улучшенного качества https://swseedco.com/press-release/calyxt-and-sws-gene-edited-alfalfa-plant-designated-as-non-regulated-by-usda/	Calyxt, TALEN	Нерегулируемый статус USDA в 2017 г.
Устойчивая к снежной плесени пшеница https://www.nature.com/news/gene-editing-surges-as-us-rethinks-regulations-1.19724	Calyxt, TALEN	
Картофель не чернеющий https://geneticliteracyproject.org/2016/10/27/calyxts-bruise-resistant-non-browning-gmo-potato-cleared-sale/	Calyxt, TALEN	Одобрен USDA в 2016 г.

Таблица 2. Окончание

Культура	Компания-разработчик, метод	**Статус**
Кукуруза с высоким содержанием крахмала (восковая кукуруза) https://www.washingtonpost.com/news/wonk/wp/2017/06/13/how-one-company-plans-to-change-your-mind-about-genetically-edited-food/	DuPont, CRISPR	Нерегулируемый статус USDA в 2016 г.
Устойчивая к засухе кукуруза https://onlinelibrary.wiley.com/doi/full/10.1111/pbi.12603	DuPont , CRISPR	Разработана в 2016 г.
Шампиньоны https://www.scientificamerican.com/article/gene-edited-crispr-mushroom-escapes-u-s-regulation/	Университет штата Пенсильвания, США, CRISPR	Нерегулируемый статус USDA в 2016 г.

ностям, связанным с новейшими технологиями, необходимо противопоставить качественно новый подход – "Безопасное проектирование". Под этим термином имеется ввиду процесс, который можно определить как внедрение процедуры идентификации и оценки рисков на ранних этапах процесса проектирования, что позволит устранить или минимизировать риски в течение всего срока жизни создаваемого организма. Предлагаемый подход состоит во внедрении технической экспертизы в максимально ранний период планирования/проектирования/разработки создаваемой растительной культуры. Интересна в этом плане практика Аргентины, которая предлагает разработчикам консультации в "Офисе регулятора" на стадии планирования [13]. Превентивность оценки рисков новых редактированных растительных культур на основании накопленных достоверных научных данных, включая многолетний мониторинг и полевые испытания, сопоставима с оценкой безопасности в авиационной или ядерной промышленности, закладывает основу "культуры безопасности" в области биотехнологии растений.

Можно сформулировать ряд задач, требующих решения в связи с внедрением этого предложения.

1. Разработать методики детекции генно-инженерных манипуляций в геноме не трансгенных биотехнологических растительных культур, которые могли бы подтвердить отсутствие нецелевых мутаций, отсутствие экспрессии нового белка (белков), обладающих аллергенностью и/или токсичностью за счет сдвига рамки считывания.

2. Показать, могут ли новые биотехнологические растения представлять какие-либо типы рисков, в частности, иные типы рисков по сравнению с уже изученными трансгенными растениями.

3. Определить существуют ли новые области использования, в которых хорошо или недостаточно понятны новые риски.

4. Разработать модель ранней оценки рисков редактированных растений, используя принцип "Безопасного проектирования". Повышение уровня безопасности может быть достигнуто за счет включения параметров/функций безопасности в стандарт разработки протоколов по сертификации безопасности.

Идентификация в геноме сельскохозяйственной продукции разнообразных искусственных вставок фрагментов ДНК. В разделе будут рассмотрены возможности биоинформатики для идентификации искусственных перестроек в геноме. За последние 30 лет биоинформатика разработала разнообразные компьютерные методы для изучения последовательностей оснований ДНК и РНК. Самой первой задачей было парное сравнение двух последовательностей оснований ДНК, РНК или двух аминокислотных последовательностей. Эта задача в настоящее время имеет наиболее полное решение, которое получено с использованием динамического программирования [14]. В этом случае имеются две последовательности и нужно сделать вывод об их подобии в условиях замены нуклеотидов или аминокислот, а также их вставок или делеций в заранее неизвестных местах и неизвестной длины. В настоящее время разработаны методы глобального и локального сравнения последовательностей [15, 16]. При глобальном выравнивании две последовательности сравниваются от начала до конца. При локальном выравнивании происходит поиск фрагментов двух последовательностей, которые совпадают наилучшим образом. Разработаны также эвристические программы семейства Blast и программа Fasta [15, 17–19]. Несмотря на то, что эти программы используют эвристические алгоритмы, они достаточно точно могут находить парное подобие между аминокислотными последовательностями.

В дальнейшем эта задача была расширена для сравнения разнообразных геномов. Созданы так называемые геномные браузеры, которые позволяют сравнивать уже не сравнительно короткие последовательности, а полные геномы между собой. К самым популярным из них можно отнести UCSC Genome Browser и Ensembl Genome Browser

[20, 21]. Разработаны также специализированные программы для сравнения последовательностей полных геномов [22, 23]. Эти инструменты позволяют сравнить любой растительный или животный геном рассматриваемой продукции с тем геномом, который уже был секвенирован. Кроме того, можно провести множественное сравнение разнообразных геномов или определенных их участков. Такое сравнение геномов позволяет сравнительно просто обнаружить вставки или делеции фрагментов ДНК, которые есть в одном геноме и которых нет в другом геноме. Это позволяет четко выявлять также и точечные мутации (SNP).

Одновременно с развитием методов сравнения последовательностей оснований ДНК интенсивно развивались банки данных, которые содержат большую часть того, что было секвенировано в мире. Можно выделить два основных банка данных. Это банк данных EMBL [24], который был создан Европейской молекулярно-биологической лабораторией (образована 20 странами-участницами и страной-партнером Австралией) и Genbank [25]. В этих банках данных собраны не только последовательности различных фрагментов ДНК, но и последовательности разнообразных полных геномов. К этим геномам относятся геномы многих бактерий, вирусов, растений и животных. Созданы также банки данных, описывающие в деталях геномы отдельных организмов, например, видов пасленовых, среди которых важные для сельского хозяйства культуры томат *Solanum lycopersicum,* картофель *S. tuberosum*, перец *Capsicum annuum* (https://www.solgenomics.net/), включая дикорастущие родственные виды данных культур, которые являются донорами различных агрономически ценных признаков (например, устойчивость к абиотическим и биотическим стрессам) при селекции сортов.

Наличие такой информации может позволить найти для сельскохозяйственного растения (СР) референсный геном. Под референсным геномом будем понимать геном традиционного СР, который был уже ранее секвенирован и внесен в базу данных, поэтому далее можно рассмотреть две ситуации: ситуацию, когда присутствует референсный геном, и ситуацию, когда референсный геном для СР отсутствует.

Референсный геном присутствует. В этом случае, если будет определена последовательность ДНК генома СР, то можно сравнить его биоинформационными методами с референсным геномом. В результате этого сравнения можно выявить разнообразные вставки или делеции фрагментов ДНК, которые могли бы быть сделаны в геноме СР, а также точечные замены оснований в геноме. Таким образом, при наличии референсного генома и секвенированного генома СР, проблема поиска каких-либо незаявленных вставок или делеций на сегодняшний день может быть решена. В этом случае можно обеспечить полный контроль за биобезопасностью СР на уровне генома.

Возникает также вопрос, какие виды вставок или делеций в геноме СР обладают наибольшей биологической опасностью. К наиболее потенциально опасным можно отнести некоторые вставки промоторных последовательностей, вставки потенциально опасных генов или вставки каких-либо вирусных последовательностей или их частей. При вставке промоторных последовательностей возможно изменение профиля экспрессии каких-либо генов в геноме СР, которое может привести к изменениям в процессах развития растения. В силу этого СР может получить новые биологические свойства, которых не было у растения с референсным геномом. Если вставляются какие-либо гены, то потенциальная опасность зависит от функционального значения введенного гена. Наиболее опасными являются гены, кодирующие разнообразные токсины. К не менее опасным вставкам в геном СР можно отнести и фрагменты ДНК, принадлежащие разнообразным вирусам как человека, так и сельскохозяйственных животных или растений. При этом даже вирусы растений не могут быть признаны полностью безопасными для человека и животных [26]. В этом случае использование таких СР может иметь серьезные последствия для населения или для сельского хозяйства РФ.

Референсный геном отсуствует. Далеко не всегда геном СР имеет референсный геном. В этом случае алгоритмы парного и множественного сравнения геномов не позволят выявить сделанные искусственно перестройки генома нового продукта сельского хозяйства, так как сравнивать геном СР не с чем, поэтому кроме сравнительных методов анализа необходимо развивать математические методы и алгоритмы аннотации потенциально опасных геномных последовательностей. Под аннотацией понимается определение функциональной роли различных последовательностей СР. В геноме СР можно проводить поиск последовательностей по их функциональному значению. Для такого поиска необходимо создать множества (базу данных) биологически опасных последовательностей ДНК, которые не должны присутствовать в геноме СР. К таким последовательностям минимально можно отнести:

1. разнообразные вирусные последовательности как человека, так и сельскохозяйственных растений или животных;

2. последовательности генов, которые кодируют бактериальные токсины или токсины любого другого происхождения;

3. все промоторные последовательности, которые могут быть найдены около генов, которые перечислены в пункте 2.

Поиск промоторных последовательностей в геномах СР важен по двум причинам. Во-первых, обнаружение промоторных последовательностей указывает на местоположение возможных генов, что может помочь в их идентификации. Во-вторых, промоторные последовательности позволяют в ряде случаев отличить гены от псевдогенов и указывают на точки начала транскрипции (**TSS**). Это означает, что идентификация промоторных последовательностей может помочь выделить те гены, которые могут транскрибироваться в СР. Для такой идентификации нужна последовательность полного генома сельскохозяйственного растения. В настоящее время полногеномное секвенирование является достаточно финансово затратной технологией, однако стоимость секвенирования падает, и в ближайшие годы станет возможным его массовое применение.

Возникает вопрос – как можно найти последовательности, перечисленные в пунктах 1–3 в геноме СР, если референсный геном отсутствует? Для поиска таких последовательностей можно использовать стандартный подход, который состоит в том, чтобы создать множество последовательностей ДНК (*M*), которые выполняют одну и ту же биологическую функцию. Это так называемая обучающая выборка. В обучающую выборку желательно включать такие подобия, которые приходятся только на потенциально биологически опасные последовательности. Таким образом, необходимо проанализировать все существующие банки данных нуклеотидных последовательностей и создать для каждой потенциально опасной последовательности из пунктов 1–3 свое собственное множество *M*.

Затем для каждого множества *M* создается скрытая марковская модель (**СММ**), которая используется при поиске последовательностей ДНК, выполняющих те же биологические функции, что и последовательности множества из множества *M* [27–29]. Эта скрытая марковская модель “сканируется” по всему геному СР и таким способом находятся последовательности, которые являются членами множества *M*. Это означает, что в геноме СР можно найти значительную часть потенциально опасных последовательностей, для которых было создано множество *M*. Таким способом были аннотированы многие гены и белковые семейства. Создать множества разнообразных патогенов, вирусных последовательностей сравнительно легко, если использовать существующие базы данных EMBL и Genbank. Однако на этом пути есть трудности, которые невозможно решить существующими математическими методами.

Проблема состоит в том, что использование СММ прекрасно работает, пока последовательности не накопили достаточное количество вставок и замен нуклеотидов, а также вставок или делеций как отдельных нуклеотидов, так и протяженных последовательностей. Если же во множестве *M* число мутаций на нуклеотид между любой парой последовательностей будет больше чем 2.4 [30], то статистически значимое множественное выравнивание последовательностей построить не удастся. Это приведет к тому, что поиск последовательностей в геноме СР, которые имеют такое же функциональное значение, как и последовательности из множества *M*, будет невозможен, то есть, последовательность с данной биологической функцией, например, потенциальный токсин, будет существовать в геноме СР, однако выявить ее существующими подходами не удастся.

Это приводит к необходимости разработки новых математических методов для идентификации разнообразных последовательностей ДНК в геномах СР, которые не имеют референсного генома. Это означает, что полная биологическая безопасность сельскохозяйственных растений, которые не имеют референсного генома, в настоящее время достигнута быть не может.

Для решения задачи полного выявления возможных вставок или делеций в геноме СР предлагается использовать новый метод для построения множественного выравнивания для сильно различающихся нуклеотидных последовательностей (MAHDS). На сайте http://victoria.biengi.ac.ru/mahds/auth этим методом можно построить множественное выравнивание для нуклеотидных последовательностей. Под сильно различающимися последовательностями будем понимать последовательности, накопившие более 2.5 случайных замен (x) на один нуклеотид относительно друг друга. MAHDS позволяет строить статистически значимые выравнивания для x в интервале от 2.4 до 4.4 (http://victoria.biengi.ac.ru/mahds/auth) [31]. Показано [31], что ранее разработанные алгоритмы могут строить статистически значимые множественные выравнивания до $x < 2.4$.

В настоящее время MAHDS был применен для построения множественного выравнивания промоторных последовательностей из генома *Arabidopsis Thaliana* и генома человека [32]. Эта работа показала, что статистически значимое множественное выравнивание для промоторных последовательностей невозможно рассчитать существующими методами, так как для них $x = 3.6$ [31]. Множественное выравнивание для 4220 промоторных последовательностей из генома риса было построено методом MAHDS. Разработан также метод создания классов промоторов на основе проведенного множественного выравнивания. Всего удалось создать 5 классов промоторных последовательностей с объемом классов более 100 промоторов. Полученные классы промоторных последовательностей были использованы для поиска других промоторных последовательностей в геноме риса. Для

каждого класса была создана профильная матрица размером (16.600) [33, 34]. Поиск потенциальных промоторных последовательностей проводился для каждой матрицы с использованием глобального выравнивания. Всего удалось выявить 145277 потенциальных промоторов. Из них 18563 приходились на промоторы известных генов, что составило около 46% от аннотированных генов. Для расчета числа ложных позитивов был применен алгоритм, разработанный для анализа случайно перемешанных нуклеотидных последовательностей полного генома риса. Число ложных позитивов в этом случае составило около 1×10^{-8} на один нуклеотид. Если в качестве контроля брали инвертированную последовательность хромосом из генома риса и применяли к ней разработанный алгоритм, то число ложных позитивов составляло 4×10^{-7}. В любом случае это значительно меньше, чем у всех используемых методов для поиска промоторных последовательностей в эукариотических геномах.

Существующие алгоритмы для предсказания промоторных последовательностей не могут построить статистически значимое выравнивание для промоторных последовательностей и поэтому они используют другие математические подходы. Это такие алгоритмы как TSSW [35], PePPER [36], G4PromFinder [37] и многие другие. Лучшие алгоритмы предсказывают ложный позитив на уровне 10^{-3}–10^{-4} на нуклеотид, в то время как геном риса содержит ~4.3×10^{8} оснований ДНК. В результате среди десятков тысяч ложных предсказаний невозможно выделить истинный промотор. Фактически, поиск промоторных последовательностей компьютерными методами в настоящее время возможен только методом MAHDS. Однако MAHDS – это только вычислительный метод, а полное подтверждение того, что выявляемые последовательности являются функциональными промоторами, возможно либо экспериментальными методами, либо путем изучения подобия найденных последовательностей с последовательностями различных транскриптом. В последнем случае можно попытаться найти TSS и провести поиск подобия последовательностей ДНК выше TSS с потенциальными промоторными последовательностями, которые обнаруживает разработанный математический метод.

Одновременно была изучена корреляция найденных потенциальных промоторных последовательностей с различными дисперсными повторами и транспозонами. Удалось показать, что ~87 тыс. промоторных последовательностей коррелирует с различными дисперсными повторами и транспозонами, которые ранее были найдены в работе [38]. 20.654 промоторных последовательностей приходятся на ранее аннотированные промоторы риса. При этом число ложных позитивов составляет не более 160 последовательностей. Остальные 37.390 потенциальных промоторных последовательностей могут представлять промоторы неизвестных генов (в частности, генов микро РНК [39]), промоторы, связанные с различными мобильными элементами генома, а также эволюционные следы расселения генов и их промоторов. Эти промоторные последовательности как раз наиболее интересны с точки зрения биобезопасности. Причина в том, что небольшое количество точечных мутаций в промоторной последовательности может перевести ген, находящийся за промоторной последовательностью, из неактивного состояния в активное, то есть может начаться активная транскрипция ранее молчащего гена. Фактически это означает, что некоторые биологические свойства сельскохозяйственного растения могут измениться, и из полностью безопасного оно может превратиться в потенциально опасное.

MAHDS-метод универсален и может быть использован для построения множественного выравнивания любых нуклеотидных последовательностей. В качестве таких последовательностей могут выступать перечисленные выше разнообразные вирусные последовательности, всевозможные гены токсинов, как животных и растений, так и человека. Необходимо только создать соответствующие множества *M* последовательностей и для них методом MAHDS построить множественные выравнивания. После этого геном СР “сканируется” каждым множеством *M* (таких множеств могут быть десятки тысяч) и выдается аргументированное заключение о присутствии потенциально опасных последовательностей, перечисленных в пунктах 1–3. После этого становится доступной быстрая идентификация возможных вставок фрагментов ДНК в геном СР даже для последовательностей, которые накопили существенно количество замен оснований и вставок или делеций.

CRISPR/Cas9-редактированные растения. Необходимо заметить, что полногеномное секвенирование и последующий биоинформационный анализ полученных данных, включая сборку генома и его аннотацию, является сегодня дорогостоящим и трудоемким подходом к сравнительной оценке геномов СР. Кроме того, недавно разработанные методы редактирования генома, такие как CRISPR/Cas9, в отличие от других известных методов (агробактериальная или баллистическая трансформация), позволяют вносить нужные модификации, не оставляя следов. Поэтому в настоящее время становится актуальной разработка новых подходов к определению возможных изменений генома искусственного происхождения. Основой для такого подхода может стать создание на базе имеющегося сегодня огромного количества информации Банка данных последовательностей, связанных с экономически ценными признаками видов и сортов агрокультур

[40, 41]. Далее, современными методами поиска [42, 43, https://crispr.cos.uni-heidelberg.de/] в отобранных последовательностях можно определить и собрать в отдельную базу данных сайты, которые с большой вероятностью могут быть использованы для дизайна так называемой гидовой РНК (определяет место CRISPR/Cas9-редактирования). Таким образом, поиск вставок, делеций, стоп-кодонов (исключающих экспрессию гена или продукцию корректного белка), несинонимичных однонуклеотидных замен может быть сужен до сравнительного анализа набора коротких последовательностей в ряде генов, ассоциированных с определенными характеристиками растений.

* * *

Полногеномное секвенирование и биоинформационные методы открывают новые уникальные возможности для оценки биобезопасности сельскохозяйственных культур. Становится возможным провести детальный анализ возможных вставок фрагментов ДНК в геноме СР и выяснить их биологическое значение. Также возможно провести быстрый скрининг СР на присутствие потенциально опасных генов, вирусных последовательностей и неспецифических промоторных последовательностей. Также можно будет почти полностью идентифицировать сельскохозяйственные растения, содержащие нежелательные или биологически опасные гены, онкогены и гены, продуктом которых являются токсины. Применение этих подходов на практике позволит значительно увеличить биобезопасность сельскохозяйственных растений.

Работа выполнена при частичной финансовой поддержке гранта РФФИ (№ 18-29-14067\18).

СПИСОК ЛИТЕРАТУРЫ

1. *Korobko I.V., Georgiev P.G., Skryabin K.G., Kirpichnikov M.P.* // Acta Naturae. 2016. V. 8. № 4. P. 6–13.
2. *Yakovleva I.V., Zhuravleva E.V., Kamionskaya A.M.* // FEBS Open Bio. 2019. V. 9. № S1. P. 280–281. https://doi.org/10.1002/2211-5463.12675
3. Guidance on Risk Assessment of Living Modified Organisms // Convention on biological diversity. 64 p. UNEP/CBD/BS/COP-MOP/6/13/Add.1 30 July 2012.
4. Методические указания МУК 4.2.1903-04. “Продукты пищевые. Метод идентификации генетически модифицированных источников (ГМИ) растительного происхождения с применением биологического микрочипа”, утвержденные Главным государственным санитарным врачом РФ 6 марта 2004 г. М.: Минздрав России, 2004. 10 с. с http://base.garant.ru/4181368/#friends; (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. M.: Ministry of Health of Russia, 2004. 10 p.)
5. Методические указания МУК 4.2.1902-04. “Определение генетически модифицированных источников (ГМИ) растительного происхождения методом полимеразной цепной реакции”, утвержденные Главным государственным санитарным врачом РФ 6 марта 2004 г. М.: Федеральный центр Госсанэпиднадзора Минздрава России, 2004. 46 с. http://base.garant.ru/4180376/; (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. M.: Federal Center for Sanitary and Epidemiological Supervision of the Ministry of Health of Russia, 2004. 46 p.)
6. *Agapito-Tenfen S.Z., Okoli A.S., Bernstein M.J., Wikmark O.G., Myhr A.I.* // Front. Plant Sci. 2018. V. 9. Article1874. P. 1–18. https://doi.org/10.3389/fpls.2018.01874
7. *Maher M.F., Nasti R.A., Vollbrecht M., Starker C.G., Matthew D.C., Voytas D.F.* // Nat. Biotechnol. 2020. V. 38. P. 84–89. doi.org/ https://doi.org/10.1038/s41587-019-0337-2
8. *Jones H.D.* // Nat. Plants. 2015. V. 1. Article 14011. https://doi.org/10.1038/nplants.2014.11
9. *Davison J., Ammann K.* // GM Crops Food. 2017. V. 8. № 1. P. 13–34. https://doi.org/10.1080/21645698.2017.1289305
10. *Metje-Sprink J., Mens J., Dodrzejewski D., Sprink T.* // Front. Plant Sci. 2019. V. 9. Article 1957. P. 133–141. https://doi.org/10.3389/fpls.2018.01957
11. *Globus R., Qimrom U.* // Cell Biochem. J. 2018. V. 119. № 2. P. 1291–1298. https://doi.org/10.1002/jcb.26303
12. *Yakovleva I.V., Vinogradova S.V., Kamionskaya A.M.* // Russian Journal of Genetics: Applied Research. 2016. V. 6. № 6. P. 646–656. https://doi.org/10.1134/S2079059716060095
13. *Whelan A.I., Lema M.A.* // GM Crops Food. 2015. V. 6. № 4. P. 253–265. https://doi.org/10.1080/21645698.2015.1114698
14. *Eddy S.R.* // Nat. Biotechnol. 2004. V. 22. № 7. P. 909–910. https://doi.org/10.1038/nbt0704-909
15. *Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J.* // J. Mol. Biol. 1990. V. 215. № 3. P. 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
16. *Needleman S.B., Wunsch C.D.* // J. Mol. Biol. 1970. V. 48. № 3. P. 443–453.
17. *Pearson W.R., Lipman D.J.* // PNAS. 1988. V. 85. № 8. P. 2444–2448. https://doi.org/10.1073/pnas.85.8.2444
18. *Mount D.W.* // CSH protocols. 2007. V. 2007. https://doi.org/10.1101/pdb.top16
19. *States D.J., Gish W., Altschul S.F.* // Methods. 1991. V. 3. № 1. P. 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., Flicek P.* // Database. 2016. V. 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., Haussler D.* // Genome Res. 2002. V. 12. № 6. P. 996–1006. https://doi.org/10.1101/gr.229102

22. *Derrien T., André C., Galibert F., Hitte C.* // Bioinformatics. 2006. V. 23. № 4. P. 498–499. https://doi.org/10.1093/bioinformatics/btl618
23. *Sinha A.U., Meller J.* // BMC Bioinformatics. 2007. V. 8. Article 82. https://doi.org/10.1186/1471-2105-8-82
24. *Hingamp P., van den Broek A.E., Stoesser G., Baker W.* // Mol. Biotechnol. 1999. V. 12. № 3. P. 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., Sayers E.W.* // Nucleic Acids Res. 2013 V. 41. № D1. P. D36–D42. https://doi.org/10.1093/nar/gks1195
26. *Nikitin N.A., Trifonova E.A., Karpova O.V., Atabekov J.G.* // Moscow Univ. Biol. Sci. Bull. 2016. V. 71. № 3. P. 128–134.
27. *Yoon B.J.* // Current Genomics. 2009. V 10. № 6. P. 402–415. https://doi.org/10.2174/138920209789177575
28. *Bateman A., Birney E., Durbin R., Eddy S.R., Finn R.D., Sonnhammer E.L.L.* // Nucleic Acids Res. 1999. V. 27. № 1. P. 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., Bateman A* // Nucleic Acids Res. 2010. V. 38. № S1. P. 211–222. https://doi.org/10.1093/nar/gkp985
30. *Korotkov E.V., Korotkova M.A.* // J. Physics: Conference Series. 2019. V. 1205. Article 012025. https://doi.org/10.1088/1742-6596/1205/1/012025
31. *Korotkov E.V., Suvorova Y.M., Kostenko D., Korotkova M.A.* // Genes. 2021. Under consideration.
32. *Korotkov E.V., Kamionskaya A.M., Korotkova M.A.* // Biotekhnologiya. 2020. V. 36. № 4. P. 7–14. https://doi.org/10.21519/0234-2758-2020-36-4-7-14
33. *Pugacheva V., Korotkov A., Korotkov E.* // Statistical Applications in Genetics and Molecular Biology. 2016. V. 26. № 5. P. 381–400. https://doi.org/10.1515/sagmb-2015-0079
34. *Suvorova Y.M., Korotkova M.A., Skryabin K.G., Korotkov E.V.* // DNA Res. 2019. V. 26. № 2. P. 157–170. https://doi.org/10.1093/dnares/dsy046
35. *Solovyev V.V., Shahmuradov I.A., Salamov A.A.* // Methods in Molecular Biology. /Ed. I. Ladunga. N.J.: Humana Pres, 2010. V. 674. P. 57–83. https://doi.org/10.1007/978-1-60761-854-6_5
36. *De Jong A., Pietersma H., Cordes M., Kuipers O.P., Kok J.* // BMC Genomics. 2012. V. 13. Article 299. https://doi.org/10.1186/1471-2164-13-299
37. *Di Salvo M., Pinatel E., Tala A., Fondi M., Peano C., Alifano P.* // BMC Bioinformatics. 2018. V. 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., Hufford M.B.* // Genome Biol. 2019. V. 20. № 4. Article 275. https://doi.org/10.1186/s13059-019-1905-y
39. *Zhou X., Ruan J., Wang G., Zhang W.* // PLoS Comput. Biol. 2007. V. 3. № 3. e37. P. 0412–0423. https://doi.org/10.1371/journal.pcbi.0030037
40. *Mohan V., Paran I.* The Capsicum Genome. Compendium of Plant Genomes // Ed. Ramchiary N., Kole C. N.Y.: Springer Cham., 2019. P. 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., Nagaraju J.* // BMC Plant Biol. 2015. V. 15. Article 207. https://doi.org/10.1186/s12870-015-0575-5
42. *Stemmer M., Thumberger T., del Sol Keyer M., Wittbrodt J., Mateo J.L.* // PLoS ONE. 2015. V. 10. № 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., Heckl D.* // Nucleic Acids Res. 2017. V. 46. № 3. P. 1375–1385. https://doi.org/10.1093/nar/gkx1268

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 of the Russian Academy of Sciences, Moscow, 119071 Russia*
**e-mail: bioinf@yandex.ru*

In Russia and in the world, the questions are acute 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 5 major transgenic crops, and producers are ready to switch to crops with an edited genome, approved in the USA, Argentina and other countries. We propose a qualitatively new approach for risks assessment of edited plants–"Safe Design", and we have also developed an extremely important fundamentally new approach for developing methods combined NGS and Bioinformatics for assessing the crop import biosafety. The proposed mathematical approach makes a reality the detailed analysis of possible insertions of DNA fragments into the genome of edited crops and a clarification of their biological significance. The developed method can be used for rapid screening of plants for the presence of potentially dangerous genes, viral sequences, and nonspecific promoter sequences.

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