14.3

Метод поиска пиков размерного стандарта при фрагментном анализе ДНК

© И.В. Заруцкий,[1,2] В.В. Манойлов,[1,2] Н.С. Самсонова,[1,3] А.И. Петров,[1] В.Е. Курочкин,[1] И.А. Леонтьев,[1] Я.И. Алексеев[4,5]

[1] Институт аналитического приборостроения РАН,
190103 Санкт-Петербург, Россия
[2] Университет ИТМО,
197101 Санкт-Петербург, Россия
[3] Физико-технический институт им. А.Ф. Иоффе РАН,
194021 Санкт-Петербург, Россия
[4] ООО „Синтол“,
127550 Москва, Россия
[5] Всероссийский научно-исследовательский институт сельскохозяйственной биотехнологии,
127550 Москва, Россия
e-mail: kolomna.88@mail.ru

(Поступило в Редакцию 27 декабря 2017 г.)

Фрагментный анализ ДНК методом капиллярного гель-электрофореза является эффективным инструментом изучения структуры ДНК, имеющим множество областей применения, в частности для генетической экспертизы сортов и сортообразцов ряда важнейших сельскохозяйственных культур. Фрагментный анализ ДНК проводится в несколько этапов. Этап обработки результатов исследования включает процедуру обнаружения пиков стандартных фрагментов ДНК. В работе перечислены особенности спектра распределения стандартных фрагментов ДНК, приводящие к необходимости разработки нового метода поиска пиков стандарта. Предложен метод их поиска, основанный на сопоставлении совокупности стандартных длин фрагментов ДНК и спектральных пиков. Описан алгоритм, реализующий этот метод. Также подробно рассмотрен критический этап алгоритма — выбор порога для процедуры обнаружения пиков в спектре. Перечислены достоинства и недостатки метода, приведены результаты тестирования.

DOI: 10.21883/JTF.2018.09.46429.2621

Введение

В молекуле дезоксирибонуклеиновой кислоты (ДНК) зашифрована информация, обусловливающая признаки организма, содержащего эту молекулу. ДНК обеспечивает хранение, передачу и реализацию генетической программы развития и функционирования живого организма, отвечает за изменчивость организма. Информация в ДНК хранится в виде кода — последовательности химически связанных между собой нуклеотидов. В каждый нуклеотид входят остатки сахара дезоксирибозы, фосфорной кислоты, гетероциклических азотистых оснований — органических соединений, состоящих из атомов водорода, углерода, азота и кислорода. В состав ДНК входят гетероциклические основания четырех типов — аденин (A), гуанин (G), цитозин (C) и тимин (T). Генетическая информация записана в определенных сочетаниях нуклеотидов, содержащих основания разного типа. Причем гетероциклические основания соединяются попарно по правилу комплементарности: аденин с тимином, гуанин с цитозином. Нуклеотиды расположены в двух одноцепочечных нитях, которые взаимодействуя между собой по правилу комплементарности, образуют двойную спираль ДНК. Размер молекул ДНК может сильно варьировать. Размер молекул двуцепочечной ДНК принято измерять в парах нуклеотидов или в парах оснований (base pair, bp) [1—-5].

Расшифровка нуклеотидной последовательности — секвенирование ДНК — позволяет идентифицировать любой организм и является „золотым стандартом“ идентификации сортов растений, пород животных, оценки и изучения генетических особенностей живых организмов. Классическим и наиболее распространенным методом секвенирования ДНК, основанным на принципе капиллярного электрофореза, является метод Сенгера [6–9].

Секвенирование ДНК

Определение порядка следования нуклеотидных звеньев в молекуле нуклеиновой кислоты стало незаменимым для фундаментальных биологических исследований, а также в таких прикладных областях деятельности, как медицинская диагностика, разработка биотехнологий, судебно-медицинская экспертиза, вирусология, биологическая систематика и др. [10,11]. Современные разработки достигли такой скорости секвенирования, которая позволяет получать полные последовательности ДНК организмов. Первые последовательности ДНК были получены в начале 1970-х годов с использованием ручных и кропотливых методов, основанных на хими-

ческой модификации ДНК с последующим селективным расщеплением, либо на ферментативном включении терминирующих дидезоксинуклеозидтрифосфатов (ддНТФ). После разработки метода на основе включения ддНТФ, содержащих отличающиеся по спектру флуоресценции красители на разных нуклеотидах, секвенирование стало проще и на порядок быстрее. Секвенирование ДНК может быть использовано для определения последовательности отдельных генов, более крупных генетических областей полных хромосом или целых геномов любого организма [12–14].

Фрагментный анализ ДНК

С помощью фрагментного анализа ДНК можно проводить следующие исследования.

• Генетическое типирование коротких тандемных повторов, которые обычно амплифицируют с использованием флуоресцентно-меченного прямого и немеченного обратного праймеров методом полимеразной цепной реакции (ПЦР). Продукты ПЦР разделяют по размеру при помощи капиллярного гель-электрофореза.

• Генотипирование с использованием полиморфизма одиночных нуклеотидов. Маркер для определения полиморфизма одиночных нуклеотидов (Single Nucleotide Polymorphism) состоит из отдельных пар оснований, варьирующих в известной последовательности ДНК, так, что формируется до 4 вариантов данного маркера.

• Фингерпринтинг. Несколько технологий на основании фрагментного анализа ДНК используют полиморфизм длин фрагментов, полученных путем ферментативной рестрикции и ПЦР для создания уникального профиля (фингерпринта), позволяющего различать образцы разной ДНК.

• Относительная флуоресценция. Данный подход сравнивает разницу в высоте пиков между двумя образцами ДНК — контрольным и исследуемым.

Фрагментный анализ ДНК состоит из следующих этапов.

• Проводят ПЦР с парами праймеров, в каждой из которых один праймер содержит на 5-м конце флуоресцентный краситель. Для одного образца методом ПЦР могут быть амплифицированы различные по размеру фрагменты ДНК, содержащие флуоресцентные красители разных цветов. Для определения длин амплифицированных фрагментов ДНК обязательно используется набор фрагментов ДНК с известными размерами (размерный стандарт), помеченных флуоресцентным красителем со спектром флуоресценции, отличным от спектра флуоресцентных красителей, которые используются для амплификации исследуемых фрагментов ДНК.

• Разделение флуоресцентно-меченных фрагментов ДНК в секвенаторе методом капиллярного электрофореза с детекцией сигнала флуоресценции, индуцированной лазером.

• Анализ результатов исследования с помощью специализированного программного обеспечения (определение размера фрагментов ДНК, определение генотипов на основании соотношения различных аллелей анализируемых маркеров).

Капиллярный гель-электрофорез ДНК

Капиллярный гель-электрофорез представляет собой процесс разделения ионизованных отрицательно заряженных фрагментов ДНК по размеру в среде полимера (геля), молекулы которого по размеру больше, либо сопоставимы с размером разделяемых фрагментов ДНК. Перед электрофорезом фрагменты ДНК вместе с другими отрицательно заряженными молекулами солей и остатками дезоксинуклеозидтрифосфатов и праймеров вводятся в капилляр с гелем методом электрокинетической инжекции. Высокое напряжение, приложенное к образцу в капилляре, приводит в движение отрицательно заряженные фрагменты ДНК. В результате электрофореза фрагменты ДНК разделяются в геле по соотношению заряд/масса. На электрофоретическую подвижность образца влияют условия проведения анализа: величина электроосмотического потока, тип буфера, концентрация, кислотность, температура капилляра, приложенное напряжение и используемый тип полимера. Достигая оптического окна вблизи положительного электрода, разделенные по размеру фрагменты ДНК пересекают лазерный луч. Лазерное излучение возбуждает флуоресценцию красителей, которыми помечены концы фрагментов ДНК. Флуоресцентное свечение разделяется на цвета дифракционной решеткой и регистрируется ПЗС камерой. Поскольку при возбуждении лазером различные флуоресцентные красители излучают свет на разных длинах волн, то все, даже совпадающие по размеру фрагменты ДНК, могут быть обнаружены, если они содержат в своем составе отличающиеся по спектру флуоресценции красители [15–17]. Сигналы флуоресценции оцифровываются, затем эти данные сохраняются в файл в формате, совместимом с программным обеспечением, используемым для анализа.

Размерный стандарт ДНК

Размерный стандарт ДНК представляет собой набор фрагментов ДНК известной длины, равномерно распределенных по всему диапазону возможных длин фрагментов исследуемого ДНК образца. Так, типичный размерный стандарт составлен из фрагментов ДНК различной длины в диапазоне от 20 до 1200 bp. Фрагменты ДНК размерного стандарта содержат на 5-м конце флуоресцентный краситель, отличный по спектру флуоресценции от спектров флуоресценции красителей, используемых для детектирования исследуемых фрагментов ДНК. Таким образом, возможно добавить размерный стандарт в каждый исследуемый образец и провести процедуру

электрофоретического разделения пробы совместно со стандартом в одном капилляре при одинаковых условиях. Определение исследуемых размеров фрагментов ДНК выполняется посредством сравнения с размерным стандартом. Поскольку вариации в условиях эксперимента приводят к тем же изменениям в подвижности как исследуемых фрагментов ДНК образца, так и фрагментов стандарта, то по положению фрагментов стандарта возможно построить калибровочную кривую размеров фрагментов ДНК для каждого образца. Калибровочная кривая дает возможность компенсировать изменения подвижности, которые могут варьировать в зависимости от капилляра и других условий эксперимента. Таким образом, сравнение длин исследуемых фрагментов ДНК с калибровочной кривой дает возможность точно определить размер каждого меченого флуоресцентным красителем фрагмента ДНК в образце.

Поиск пиков размерного стандарта ДНК

Флуоресценция красителя, входящего в размерный стандарт регистрируется в оптическом окне капилляра в виде хроматографических пиков, положение которых по оси абсцисс соответствует длинам фрагментов ДНК стандарта. Кроме пиков, соответствующих фрагментам размерного стандарта ДНК, в хроматограмме всегда присутствуют дрейф базовой линии, шум, а также „выбросы“ — артефакты процессов разделения и регистрации. Шум не представляет проблемы, поскольку в зарегистрированных спектрах отношение сигнал/шум равно 10 или более и при правильном выборе порога обнаружения не препятствует надежному детектированию пиков. Дрейф базовой линии также без затруднений компенсируется известными методами [18]. Серьезную помеху представляют „выбросы“, поскольку имеют амплитуду и ширину, сопоставимые с таковыми у пиков стандарта, в результате чего „выбросы“ регистрируются в качестве ложных пиков стандарта, что приводит к ошибкам в коррекции подвижности, что в свою очередь приводит к неправильному анализу фрагментов ДНК исследуемого образца. Таких ложных пиков бывает в спектре довольно много — до 10% от общего количества. Фильтрация „выбросов“ обычными методами [19–21] затруднительна, так как „выбросы“ неотличимы от пиков стандарта ни по амплитуде, ни по ширине, ни по форме, ни по частотному спектру. Непосредственно по времени удерживания пики стандарта от ложных отделить также невозможно, так как в результате изменения подвижности положения пиков стандарта и дистанция между ними не воспроизводятся от эксперимента к эксперименту. Таким образом, рассматривая отдельный пик, ничего нельзя сказать о том, принадлежит он стандарту или нет. Это приводит к необходимости рассматривать совокупность пиков. Поскольку длины всех фрагментов размерного стандарта ДНК известны, а зависимость времени удерживания от размера фрагмента мало отличается от линейной, то можно в качестве пиков стандарта отобрать только те, положение которых укладывается в почти линейную стандартную кривую. Проблема лишь в том, что коэффициент пропорциональности для построения линейной последовательности и положение первого пика стандарта заранее не известны, и к тому же коэффициент пропорциональности зависит от времени удерживания. Для решения этой проблемы предлагается следующая идея: рассмотреть все возможные стандартные кривые, для всех возможных коэффициентов пропорциональности, задаваемых случайными комбинациями положений двух любых пиков. Для ускорения сходимости можно использовать следующее, выполняющееся на практике, предположение: чем больше амплитуда пика — тем выше вероятность того, что пик принадлежит размерному стандарту. Такой метод поиска пиков стандарта реализуется следующим алгоритмом.

1. Оценивается уровень шума в хроматограмме и вычисляется порог для обнаружения пиков.

2. Производится поиск в хроматограмме всех пиков, амплитуды которых больше порога.

3. Из найденных пиков отбираются N наиболее вероятных претендентов (пики с максимальной амплитудой), для первой итерации N равно количеству длин фрагментов в размерном стандарте.

4. Из отобранных пиков выбираются два любых пика (например, первый и второй, или последний и предпоследний).

5. Положение первого служит началом для стандартной кривой, дистанция между пиками определяет начальное значение коэффициента пропорциональности.

6. По стандартной кривой вычисляются ожидаемые положения пиков стандарта.

7. Пики хроматограммы, находящиеся в ожидаемых положениях помечаются как принадлежащие стандарту, в ходе проверки положений пиков вносится поправка в стандартную кривую. Поправка определяется по фактическому положению пика.

8. Если количество помеченных пиков равно количеству длин фрагментов стандарта — стандарт найден, и все помеченные пики принадлежат стандарту. В противном случае выбираем следующую пару пиков и переходим к п. 5. Если все возможные пары пиков из N отобранных проверены, но стандарт не найден — увеличиваем N на единицу и переходим к п. 3.

Возможна ситуация, когда ни одна из возможных стандартных кривых не удовлетворяет п. 8, это свидетельствует о плохой подготовке пробы или сбоях при проведении хроматографического разделения, такие пробы бракуются.

Оценивание уровня шума и вычисление порога для процедуры обнаружения пиков

При поиске пиков в спектре размерного стандарта критически важно правильно выбрать порог обнаружения. Завышенный порог приведет к пропуску некоторых пиков стандарта, в результате чего стандарт не будет обнаружен. При пороге, заниженном до уровня шума, количество обнаруженных пиков и их плотность возрастает настолько, что становится возможным построить несколько стандартных кривых, для каждой из которых найдутся все пики в ожидаемых положениях. В этом случае стандарт не будет найден, поскольку нет способа выделить истинную кривую из нескольких возможных. Таким образом, выбор величины порога обнаружения должен определяться стандартным отклонением шума. Величина стандартного отклонения шума и отношение сигнал/шум зависят от подготовки пробы и условий проведения электрофореза, по этой причине невозможно определить все параметры шума до проведения эксперимента. На основе анализа работы прибора и статистического исследования о характеристиках шума известно следующее: шум стационарный, среднее значение равно нулю, плотность вероятности распределена по нормальному закону. О пиках известны их форма и средняя ширина. Оценку величины стандартного отклонения (с.к.о.) необходимо вычислять отдельно для каждой пробы на основе зарегистрированного хроматографического спектра.

Для процедуры поиска пиков стандарта авторами разработан и реализован алгоритм, позволяющий автоматически получать оценку с.к.о. шума и вычислять величину порога.

Алгоритм выполняется в четыре этапа.

1. Коррекция базовой линии. Для вычисления и вычитания базовой линии используется метод, описанный в [18]. На рис. 1 представлен сигнал спектра стандарта, который содержит полезную составляющую в виде пиков, базовую линию и шум. На рис. 2 показан спектр после коррекции базовой линии. Вся дальнейшая обработка проводится над скорректированным сигналом.

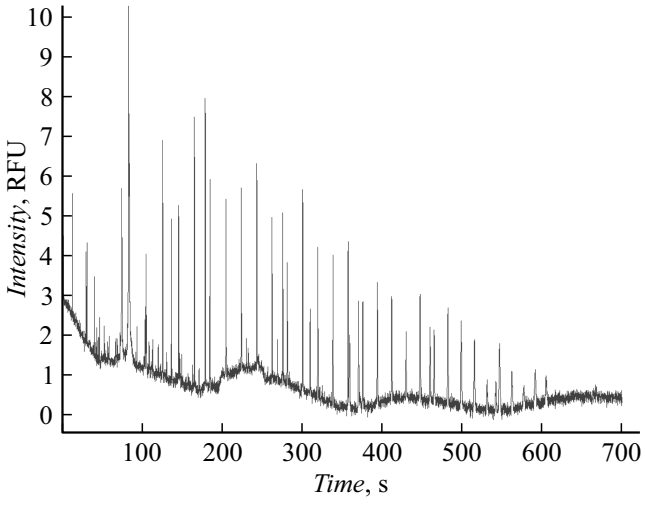

Рис. 1. Исходный спектр стандартных фрагментов ДНК.

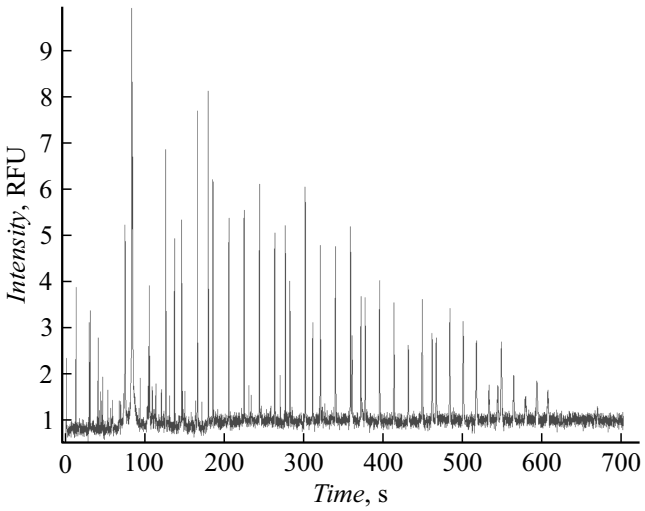

Рис. 2. Спектр после коррекции дрейфа базовой линии.

2. Вычитание из спектра пиков и выбросов шума. Как упоминалось выше, пики спектра и выбросы шума неотличимы, поэтому и те и другие обнаруживаются одинаковым образом. Для их обнаружения используется метод поиска на основе согласованной фильтрации [22]. В этом методе используется свертка исходного сигнала с сигналом, описывающим форму пика. Свертка вычисляется следующим образом:

$$s_1(t) = \int_U^T s(t) f(t-\tau) d\tau,$$

где t — независимая переменная — время, $s(t)$ — зарегистрированная хроматограмма, $f(t)$ — форма пика, T — длина хроматограммы.

Форма пика определяется аппаратной функцией прибора и с достаточной для данного алгоритма точностью описывается функцией на основе гауссиана:

$$f(t) = \exp\left[-\left(-\frac{t}{w}\right)^2\right],$$

где w — средняя полуширина пика в хроматограмме.

В результате операции свертки отношение сигнал/шум в $s_1(\tau)$ возрастает в 4–5 раз по сравнению в исходным сигналом, что позволяет надежно обнаружить пики спектра и определить их положение. Если величина $s_1(t)$ превышает порог h_1, то эту точку t считаем принадлежащей пику и отсчет спектра в этой точке заменяем нулем. Порог h_1 вычисляется следующим образом: $h_1 - 2|s_{\min}|$, где $s_{\min}$ — минимум сигнала $s(t)$. Сигнал после вычитания пиков и выбросов шума показан на рис. 3.

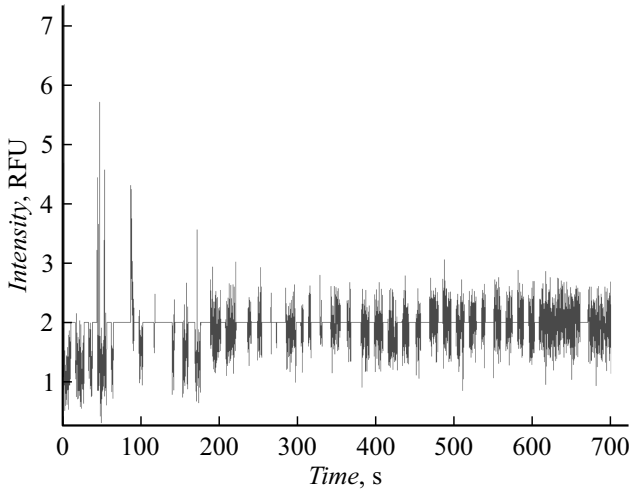

Рис. 3. Спектр после замены пиков нулями.

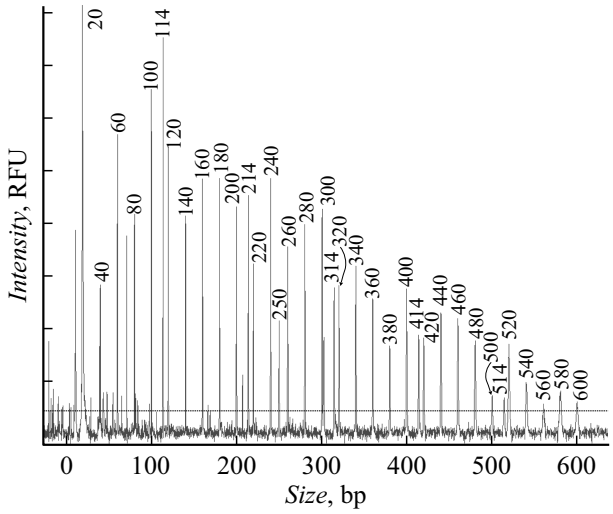

Рис. 4. Размеченный размерный стандарт, пунктирной линией отмечен уровень порога обнаружения.

3. Вычисление оценки с.к.о. шума. Определение стандартного отклонения шума производится по фрагментам между нулями. Для каждого фрагмента вычисляется с.к.о, полученные с.к.о. сортируются по величине, в качестве конечной оценки шума всего спектра берется медиана этого ряда.

4. Вычисление порога h для процедуры обнаружения пиков в спектре. Вычисление проводится следующим образом: $h = 4\sigma$, где σ — оценка с.к.о. шума, полученная на этапе 3. На рис. 4 пунктирной линией показан порог обнаружения, полученный в результате работы данного алгоритма.

Достоинства и недостатки метода

Из достоинств метода можно отметить следующее.
- Минимум априорной информации. Для работы алгоритма необходимо заранее знать среднюю полуширину пика в спектре и базовые параметры шума. Эти величины не зависят от срока службы прибора и хорошо воспроизводятся от прибора к прибору. Вся остальная необходимая информация извлекается непосредственно из зарегистрированного спектра.
- Простота реализации.
- Быстрая сходимость.

Алгоритм, реализующий метод, имеет один заметный недостаток: в спектре должны присутствовать и должны быть обнаружены все пики стандарта, если по каким-либо причинам отсутствует хотя бы один пик — размерный стандарт не будет найден.

Заключение

Предложенный метод прост в реализации и пригоден для практического применения в программном обеспечении генетического анализатора. Тестирование метода показало достаточную надежность: одна ошибка второго рода (не найден стандарт в годном спектре) и ноль ошибок первого рода (обнаружен ложный стандарт) на 10^6 успешно обработанных спектров.

Исследование выполнено при финансовой поддержке ФАНО России в рамках выполнения подпрограммы „Развитие селекции и семеноводства картофеля в Российской Федерации“ Федеральной научно-технической программы развития сельского хозяйства на 2017–2025 гг. по теме: „Разработка программного обеспечения для генетической экспертизы сортов и сортообразцов картофеля“. Номер дополнительного государственного задания Института аналитического приборостроения РАН: № 007-02-2014 от 28.11.2017 г. Для выполнения работы использовалось дорогостоящее оборудование Центра коллективного пользования „Биотехнология“ ФГБНУ ВНИИСБ — генетический анализатор „НАНОФОР 05“.

Список литературы

[1] *Watson J., Crick F.* // Nature. 1953. Vol 171. N 4356. P. 737–738.
[2] The Nobel Prize in Physiology or Medicine 1962. https://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/
[3] *Crick F.H., Barnett L., Brenner S., Watts-Tobin R.J.* // Nature. 1961. Vol. 192. N 4809. P. 1227–1232.
[4] *Овчинников Ю.А.* Биоорганическая химия. М.: Просвещение, 1987. 816 с.
[5] *Blackburn M.G.* Nucleic Acids in Chemistry And Biology. Great Britain: Royal Society of Chemistry, 2006. P. 168.
[6] *Sanger F., Nicklen S., Coulson A.R.* // Proc. Nat. Acad. Sci. USA. 1977. Vol. 74. N 12. P. 5463–5467.
[7] *Fiers W., Contreras R., Haegemann G., Rogiers R., Van de Voorde A., Van Heuverswyn H., Van Herreweghe J., Volckaert G., Ysebaert M.* // Nature. 1978. Vol. 273. N 5658. P. 113–120.
[8] *Reddy V.B., Thimmappaya B., Dhar R., Subramanian K.N., Zain B.S., Pan J., Ghosh P.K., Celma M.L., Weissman S.M.* // Science. 1978. Vol. 200. N 4341. P. 494–502.

[9] *Nunnally B.K., He H., Li L.C., Tucker S.A., McGown L.B.* // Anal. Chem. 1997. Vol. 69. N 13. P. 2392–2397.
[10] *Bergot B.J., Chakerian V., Connell C.R., Eadie J.S., Fung S., Hershey N.D., Lee L.G., Menchen S.M., Woo S.L.* Patent 5366860. US. 1989.
[11] *Lee L.G., Spurgeon S.L., Heiner C.R., Benson S.C., Rosenblum B.B., Menchen S.M., Graham R.J., Constantinescu A., Upadhya K.G., Cassel J.M.* // Nucl. Acid. Res. 1997. Vol. 25. N 14. P. 2816–2822.
[12] *Menchen S.M., Lee L.G., Connell C.R., Hershey N.D., Chakerian A., Woo S., Fung S.* Patent 5188934. US. 1993.
[13] *Rosenblum B.B., Lee L.G., Spurgeon S.L., Khan S.H., Menchen S.M., Heiner C.R., Chen S.M.* // Nucl. Acid. Res. 1997. Vol. 25. N 22. P. 4500–4504.
[14] *Ju J., Ruan C., Fuller C.W., Glazer A.N., Mathies R.A.* // Proc. Natl. Acad. Sci. USA. 1995. Vol. 92. N 10. P. 4347–4351.
[15] *Mujumdar R.B., Ernst L.A., Mujumdar S.R., Lewis C.J.* // Bioconjugate Chem. 1993. Vol. 4. N 2. P. 105–111.
[16] *Tu O., Knott T., Marsh M., Bechtol K., Harris D., Barker D., Bashkin J.* // Nucl. Acid. Res. 1998. Vol. 26. N 11. P. 2797–2802.
[17] *Tabor S., Richardson C.C.A.* // Proc. Natl. Acad. Sc.i USA. 1995. Vol. 92. N 14. P. 6339–6343.
[18] *Белов Д.А., Белов Ю.В., Манойлов В.В., Курочкин В.Е.* // Научное приборостроение. 2014. Т. 24. Вып. 3. С. 87–91.
[19] *Заруцкий И.В., Манойлов В.В.* // Научное приборостроение. 2007. Т. 17. Вып. 1. С. 115–120.
[20] *Манойлов В.В., Заруцкий И.В.* // Научное приборостроение. 2002. Т. 12. Вып. 3. С. 38–46.
[21] *Манойлов В.В., Заруцкий И.В.* // Научное приборостроение. 2009. Т. 19. Вып. 3. С. 35–40.
[22] *Кук Ч., Бернфельд М.* Радиолокационные сигналы / Пер. с англ. под ред. В.С. Кельзона. М.: Сов. радио, 1971. 568 с.