Известия Саратовского университета. Новая серия. Серия: Физика. 2023. Т. 23, вып. 1. С. 46–55
Izvestiya of Saratov University. Physics, 2023, vol. 23, iss. 1, pp. 46–55
https://fizika.sgu.ru https://doi.org/10.18500/1817-3020-2023-23-1-46-55, EDN: IQKRQK

Научная статья
УДК 535.512

Малоугловая поляриметрия как метод идентификации последовательностей нуклеотидов в биоинформатике

Д. А. Зимняков[1,2], **М. В. Алонова**[1]✉, **А. В. Скрипаль**[3], **С. Ю. Добдин**[3], **В. А. Федорова**[3]

[1]Саратовский государственный технический университет имени Гагарина Ю. А., Россия, 410054, г. Саратов, ул. Политехническая, д. 77
[2]Институт проблем точной механики и управления Российской академии наук, Россия, 410028, г. Саратов, ул. Рабочая, д. 24
[3]Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, Россия, 410012, г. Саратов, ул. Астраханская, д. 83

Зимняков Дмитрий Александрович, доктор физико-математических наук, профессор, [1]заведующий кафедрой «Физика»; [2]главный научный сотрудник лаборатории проблем лазерной диагностики технических и живых систем, zimnykov@mail.ru, https://orcid.org/0000-0002-9787-7903

Алонова Марина Васильевна, кандидат физико-математических наук, доцент кафедры «Физика», alonova_marina@mail.ru, https://orcid.org/0000-0001-7772-3985

Скрипаль Анатолий Владимирович, доктор физико-математических наук, профессор, заведующий кафедрой медицинской физики, skripalav@info.sgu.ru, https://orcid.org/0000-0002-9080-0057

Добдин Сергей Юрьевич, кандидат физико-математических наук, доцент кафедры физики твёрдого тела, dobdinsy@info.sgu.ru, https://orcid.org/0000-0002-0801-4664

Федорова Валентина Анатольевна, доктор медицинских наук, профессор кафедры медицинской физики, feodorovav@mail.ru, https://orcid.org/0000-0002-3827-407X

Аннотация. Рассмотрен метод идентификации символьных последовательностей, ассоциируемых с генетической структурой биологических объектов, с использованием принципов малоугловой поляриметрии. В рамках метода анализируемая символьная последовательность представляется двумерной фазомодулирующей матрицей, каждый элемент которой соответствует одному из четырех базовых нуклеотидов (аденину, цитозину, тимину, гуанину), а глубина модуляции фазы считывающего когерентного линейно поляризованного пучка определяется содержанием данного нуклеотида в соответствующем триплете в последовательности нуклеотидов. В результате дифракции считывающего когерентного пучка с плоскостью поляризации, ориентированной под углом 45° к сторонам фазомодулирующей матрицы, в приосевой области дальней зоны дифракции формируется пространственное распределение локальных состояний поляризации дифрагировавшего на матрице считывающего поля. Дискриминация локальных состояний поляризации в соответствии с предложенным алгоритмом позволяет синтезировать бинарное пространственное распределение, являющееся уникальным идентификатором анализируемой символьной последовательности. Моделирование процессов фазового кодирования и последующего анализа локальных состояний поляризации в приосевой области с использованием результатов секвенирования для штаммов «Ухань», «Дельта» и «Омикрон» вируса SARS-CoV-2 показало высокую чувствительность метода к локальным изменениям в структуре последовательностей нуклеотидов.

Ключевые слова: генетические структуры, последовательности нуклеотидов, фазовое кодирование, малоугловая поляриметрия, компоненты вектора Стокса

Благодарности: Работа выполнена при финансовой поддержке Российского научного фонда (грант РНФ № 22-21-00194).

Для цитирования: *Зимняков Д. А., Алонова М. В., Скрипаль А. В., Добдин С. Ю., Федорова В. А.* Малоугловая поляриметрия как метод идентификации последовательностей нуклеотидов в биоинформатике // Известия Саратовского университета. Новая серия. Серия: Физика. 2023. Т. 23, вып. 1. С. 46–55. https://doi.org/10.18500/1817-3020-2023-23-1-46-55, EDN: IQKRQK

Article

Small-angle polarimetry as a technique for identification of nucleotide sequences in bioinformatics

D. A. Zimnyakov[1,2], **M. V. Alonova**[1]✉, **A. V. Skripal**[3], **S. Yu. Dobdin**[3], **V. A. Feodorova**[3]

[1]Yury Gagarin State Technical University of Saratov, 77 Polytechnicheskaya St., Saratov 410054, Russia
[2]Institute for Precision Mechanics and Control Problems of the Russian Academy of Sciences, 24 Rabochaya St., Saratov 410028, Russia
[3]Saratov State University, 83 Astrakhanskaya St., Saratov 410012, Russia

Dmitry A. Zimnyakov, zimnykov@mail.ru, https://orcid.org/0000-0002-9787-7903
Marina V. Alonova, alonova_marina@mail.ru, https://orcid.org/0000-0001-7772-3985
Anatoly V. Skripal, skripalav@info.sgu.ru, https://orcid.org/0000-0002-9080-0057
Sergey Yu. Dobdin, dobdinsy@info.sgu.ru, https://orcid.org/0000-0002-0801-4664
Valentina A. Feodorova, feodorovav@mail.ru, https://orcid.org/0000-0002-3827-407X

Abstract. ***Background and Objectives***: The method of identification of symbolic sequences associated with the genetic structure of biological objects using the principles of small-angle polarimetry is considered. This method of analyzing and visualizing symbolic sequences obtained by sequencing DNA fragments can be defined as small-angle polarimetry of phase-modulating structures associated with genetic information. ***Materials and Methods*:** The analyzed symbolic sequence is represented by a two-dimensional phase-modulating matrix, each element of which corresponds to one of the four basic nucleotides (adenine, cytosine, thymine, guanine), and the depth of modulation of the phase of the reading coherent linearly polarized beam is determined by the content of this nucleotide in the corresponding triplet in the nucleotide sequence. As a result of the diffraction of a reading coherent beam with a polarization plane oriented at an angle of 45° to the sides of the phase-modulating matrix, a spatial distribution of local polarization states of the reading field diffracted on the matrix is formed in the paraxial region of the far diffraction zone. Discrimination of local polarization states in accordance with the proposed algorithm makes it possible to synthesize a binary spatial distribution, which is a unique identifier of the analyzed symbol sequence. ***Results*:** Modeling of the processes of phase coding and subsequent analysis of local polarization states in the near-axial region using sequencing results for the strains "Wuhan", "Delta" and "Omicron" of the SARS-CoV-2 virus has shown a high sensitivity of the method to local changes in the structure of nucleotide sequences. ***Conclusion*:** The results of the simulation allow us to conclude that binary distributions of local polarization states of light fields diffracted on DNA-associated phase-modulating structures recorded in the axial region are characterized by high sensitivity to local mutational changes in the structure of nucleotide sequences. The results obtained can be used as a basis for creating effective hybrid methods for analyzing genetic information using the principles of polarization coding and small-angle polarimetry.

Keywords: genetic structures, nucleotide sequences, phase coding, small-angle polarimetry, components of the Stokes vector

Acknowledgments: This work was supported by the Russian Science Foundation (project no. 22-21-00194).

For citation: Zimnyakov D. A., Alonova M. V., Skripal A. V., Dobdin S. Yu., Feodorova V. A. Small-angle polarimetry as a technique for identification of nucleotide sequences in bioinformatics. *Izvestiya of Saratov University. Physics*, 2023, vol. 23, iss. 1, pp. 46–55 (in Russian). https://doi.org/10.18500/1817-3020-2023-23-1-46-55, EDN: IQKRQK

Введение

Развитие новых методов анализа генетической информации, получаемой в результате секвенирования ДНК и РНК различных биологических объектов, является одним из ключевых направлений в современной биоинформатике. В значительной степени это обусловлено участившимися в последние годы случаями появления новых штаммов патогенных микроорганизмов, характеризующихся высокой степенью изменчивости и приводящих к массовым заболеваниям людей и животных с высокой вероятностью летального исхода. Внедрение в практику эффективных технологий секвенирования фрагментов ДНК и РНК стимулировало, в свою очередь, развитие биоинформационных технологий, основанных на различных подходах к анализу и идентификации генетической информации, представляемой в форме символьных последовательностей вида ATCTTGAAT… Каждый элемент в подобных последовательностях ассоциируется с одним из четырех базовых нуклеотидов (A – аденин, T – тимин, C – цитозин, G – гуанин), а структура последовательности, определяемая порядком расположения в ней нуклеотидов, уникальным образом определяет исследуемый биологический объект. В течение последних пятидесяти лет в биоинформатике сформировалось направление, основанное на компьютерном анализе ассоциируемых с генетическими структурами символьных последовательностей с использованием различных методов статистического и корреляционного анализа, нейросетевых технологий, элементов искусственного интеллекта и др. [1–10]. Эти методы, в частности, реализованы в таких широко используемых специальных программах, как BLAST (Basic Local Alignment Search Tool), GRAIL, GENSCAN, HEXON, FGENESH и т. д. (см., например, [11–14]).

Наряду с чисто программными методами анализа последовательностей нуклеотидов, для решения задач биоинформатики могут быть также применены гибридные инструментально-программные подходы, в которых реализованы принципы когерентно-оптической обработки и преобразования информации. При этом исходные символьные последовательности трансформируются в двумерные фазомодулирующие структуры (например, путем использования жидкокристаллических пространственных модуляторов света), которые затем считываются когерентным световым пучком. Дифракция считывающего пучка на фазомодулирующей структуре приводит

к формированию в дальней зоне дифракции (например, в фокальной плоскости фурье-преобразующей линзы) сложного пространственного распределения комплексной амплитуды светового поля, уникальным образом связанного с фазомодулирующей структурой (и, соответственно, с исходной символьной последовательностью). Таким образом, в оптическом блоке подобного гибридного анализатора производится частичная квазипараллельная обработка кодированных с помощью пространственного модулятора исходных данных, сводящаяся к двумерному фурье-преобразованию пространственного распределения фазы светового поля, кодирующего исходную символьную последовательность. Отметим, что обсуждаемый гибридный подход имеет отношение не к технологии секвенирования фрагментов ДНК различных биологических объектов (т. е. к процессу первичного получения информации о генетической структуре), а к последующему анализу генерируемых на первом этапе символьных последовательностей, характеризующих распределения базовых нуклеотидов во фрагментах ДНК. Число символов в таких последовательностях, как правило, составляет несколько тысяч. Соответственно, синтезируемые фазомодулирующие матрицы содержат порядка нескольких десятков столбцов и строк и могут быть физически реализованы путем использования коммерчески доступных жидкокристаллических пространственных модуляторов света. В качестве примера можно рассмотреть жидкокристаллические модуляторы, выпускаемые фирмой Holoeye Photonics AG (Германия), в частности, модулятор SLM LC 2012 с размером пикселей 36 мкм, рабочей зоной 36.9×27.6 мм и восьмибитным представлением вводимой информации вполне пригоден для синтеза фазомодулирующих структур под компьютерным управлением. Считывание кодированной информации может осуществляться пучком одномодового гелий-неонового лазера с линейной поляризацией, а регистрация дифрагировавшего светового поля может осуществляться КМОП-камерой в фокальной плоскости фурье-преобразующей линзы.

Следует также отметить, что возможности когерентно-оптического процессора как составной части гибридного анализатора не ограничиваются только прямым и обратным фурье-преобразованием кодированной информации; возможна реализация других типов линейных интегральных преобразований (например, Гильберта, Меллина и др. [15]), а также осуществление операций двумерной корреляции и свертки.

Может показаться, что обсуждаемый гибридный подход при наличии достаточно развитого набора специального программного обеспечения является избыточным и более сложным в реализации вследствие дополнительной инструментальной составляющей. Тем не менее, его использование представляет определенный интерес не только с точки зрения анализа структуры последовательностей нуклеотидов, но также и с точки зрения их визуального отображения. Кроме того, физические принципы, лежащие в основе когерентно-оптического анализа синтезируемых фазомодулирующих структур, могут быть применены при создании новых эффективных алгоритмов компьютерной обработки данных в биоинформатике.

В [16, 17] предложен возможный вариант гибридного подхода, использующий принцип формирования и анализа так называемых GB (gene-based) спеклов. При этом анализируемой последовательности нуклеотидов ставится в соответствие двумерная фазомодулирующая структура (фазовый экран), каждый элемент которой ассоциируется с определенным триплетом (комбинацией из трех базовых нуклеотидов) в последовательности. Синтезируемый фазовый экран состоит из $N \times N$ элементов; таким образом, число триплетов в последовательности равно N^2. При кодировании, начиная со старт-кодона, выбирается фрагмент последовательности с числом триплетов, равным максимально возможному квадрату целого числа. Числовые значения элементов синтезируемой квадратной матрицы изменяются в пределах от 0 до 63 и определяются следующим правилом кодирования:

$$X_{i,j} = 16\Xi_1 + 4\Xi_2 + \Xi_3 - 21. \qquad (1)$$

В выражении (1) каждый из трех весовых коэффициентов Ξ_1, Ξ_2, Ξ_3 принимает значения от 1 до 4 в соответствии со следующими ассоциациями: $A \leftrightarrow 1$, $C \leftrightarrow 2$, $G \leftrightarrow 3$, $T \leftrightarrow 4$. Нижние индексы 1–3 определяют положение нуклеотида в триплете. Таким образом, минимально возможное значение $X_{i,j}$, равное 0, соответствует триплету AAA, а максимальное значение, равное 63, достигается для триплета TTT. Представленный выше выбор ассоциаций между весовыми коэффициентами Ξ_{13} и базовыми нуклеотидами произволен и может быть заменен любым другим, например $A \leftrightarrow 4$, $C \leftrightarrow 3$, $G \leftrightarrow 2$, $T \leftrightarrow 1$.

Сформированная подобным образом матрица интерпретируется как фазовый экран, осуществляющий модуляцию фазы проходящего через него считывающего когерентного пучка в соответствии с правилом $\Delta\phi_{i,j} = K_\varphi \cdot X_{i,j}$, где K_φ – фактор модуляции фазы. Формируемое в дальней зоне дифракции спекл-модулированное распределение интенсивности светового поля является уникальным идентификатором анализируемой последовательности нуклеотидов. Соответственно, какие-либо мутационные изменения в структуре последовательности (связанные, например, с замещением части нуклеотидов) приводят к изменениям в формируемой спекл-структуре. В [16, 17] предложено осуществлять идентификацию сходства и различия в анализируемых последовательностях с применением принципов когерентно-оптического распознавания образов на основе согласованной фильтрации [15]. Однако предварительный анализ показывает, что малые изменения в структурах анализируемых последовательностей, связанные с замещением одного – трех нуклеотидов, приводят к весьма несущественной декорреляции спекл-структур для измененных последовательностей по отношению к спекл-структурам, соответствующим референтным последовательностям. Другим фактором, существенно ограничивающим практическое использование когерентно-оптического распознавания GB спекл-структур в соответствии с предложенным в [16, 17] подходом, является сложность инструментальной реализации когерентно-оптической системы распознавания образов и высокие требования к юстировке и механической стабильности элементов подобной системы. Наконец, характерной особенностью формируемых GB спекл-структур является высокая пространственная неоднородность распределений интенсивности, обусловленная весьма существенным вкладом недифрагировавшей составляющей прошедшего через фазовый экран считывающего пучка в приосевой области зоны дифракции. Наличие недифрагировавшей («когерентной») составляющей даже при больших значениях K_φ затрудняет процедуру считывания и анализа GB спекл-структур вследствие значительного динамического диапазона значений интенсивности в считываемой структуре.

В данной работе рассмотрен альтернативный подход к когерентно-оптическому анализу последовательностей нуклеотидов, основанный на применении принципов поляризационного кодирования последовательностей нуклеотидов и считывания локальных состояний поляризации дифрагировавшего пучка в приосевой области зоны дифракции, характеризуемой большими значениями интенсивности. Этот подход свободен от упомянутых выше недостатков метода анализа GB спеклов. Кроме того, как показано ниже, он характеризуется достаточно высокой чувствительностью к локальным изменениям в структуре анализируемых последовательностей. Обсуждаемая методика анализа и визуализации символьных последовательностей, получаемых в результате секвенирования фрагментов ДНК, может быть определена как малоугловая поляриметрия фазомодулирующих структур, ассоциируемых с генетической информацией.

Физические принципы малоугловой поляриметрии ДНК-ассоциированных символьных последовательностей

Предлагаемая методика малоугловой поляриметрии состоит из следующих стадий:

1) синтез двумерного фазового экрана на основе анализируемой ДНК-ассоциированной символьной последовательности;

2) формирование поляризационно-зависимой дифракционной структуры в приосевой области дальней зоны дифракции путем считывания синтезированного экрана коллимированным линейно поляризованным когерентным пучком и последующего фурье-преобразования распределения фазы в прошедшем через экран пучке с использованием собирающей линзы;

3) выделение области дальней зоны дифракции, для которой локальные состояния поляризации дифрагировавшего пучка удовлетворяют заданному критерию дискриминации.

Полученное на завершающей стадии бинарное распределение локальных состояний, удовлетворяющих критерию дискриминации, может рассматриваться как уникальный идентификатор анализируемой символьной последовательности.

В отличие от рассмотренного в [16, 17] метода кодирования ДНК-ассоциированных символьных последовательностей (выражение (1)), в рассматриваемом случае каждый триплет в последовательности нуклеотидов представляется субматрицей размера (2×2), каждому элементу которой ставится в соответствие один из четырех базовых нуклеотидов (например, $\tilde{a}_{11} \to A$, $\tilde{a}_{12} \to C$, $\tilde{a}_{21} \to T$, $\tilde{a}_{22} \to G$), а значения элементов определяют число соответствующих нуклеоти-

дов в триплете. Например:

$$\begin{pmatrix} 2 & 0 \\ 1 & 0 \end{pmatrix} \rightarrow AAT, \text{ или } TAA, \text{ или } ATA. \quad (2)$$

Отметим, что подобный подход к кодированию чувствителен к числу различных нуклеотидов в триплете, но не к их взаимным позициям в пределах триплета. Тем не менее, он позволяет успешно решать, например, задачи идентификации символьных последовательностей, их частотного анализа и др. После представления триплетов нуклеотидов набором субматриц формируется основная фазомодулирующая матрица $(a_{ik})_{2N\times 2N}$ путем построчной сборки субматриц в соответствии с порядком расположения триплетов в анализируемой последовательности. Очевидно, что размер анализируемой последовательности должен быть равен N^2.

Предположим, что синтезированный фазовый экран считывается коллимированным линейно поляризованным пучком с плоскостью поляризации, образующей угол $\frac{\pi}{4}$ со сторонами экрана, а фазовая модуляция x- и y-поляризованных составляющих пучка осуществляется в соответствии со следующим правилом:

$$\begin{aligned} (\Delta\phi_{ij})^{x}_{2N\times 2N} &= 0, \\ (\Delta\phi_{ij})^{y}_{2N\times 2N} &= \left(\frac{\pi}{2}\right)\cdot(a_{ij})_{2n\times 2n}. \end{aligned} \quad (3)$$

Распределение амплитуд ортогонально поляризованных составляющих дифрагировавшего поля в дальней зоне дифракции может быть описано следующим выражением [18, 19]:

$$E^{x,y}_{k,m} = \frac{1}{4N^2}\sum_{i=-N}^{N-1}\sum_{j=-N}^{N-1}\exp\left[-\tilde{j}\cdot K_{SC}\times \right. \\ \left. \times\left\{\left(\frac{\pi}{N}\right)(k\cdot i + m\cdot j) - \Delta^{x,y}_{\varphi i,j}\right\}\right]. \quad (4)$$

Здесь $\tilde{j}$ – мнимая единица, K_{sc} – масштабный фактор, определяющий размеры анализируемой области в пределах дальней зоны дифракции (k,m), а локальные состояния поляризации в пределах анализируемой области описываются нормированными компонентами вектора Стокса [20]:

$$\begin{cases} s^0_{k,m} = \left(\left|E^x_{k,m}\right|^2 + \left|E^y_{k,m}\right|^2\right)\Big/ 2; \\ s^1_{k,m} = \left(\left|E^x_{k,m}\right|^2 - \left|E^y_{k,m}\right|^2\right)\Big/ 2s^0_{k,m}; \\ s^2_{k,m} = 2\left|E^x_{k,m}\right|\left|E^y_{k,m}\right|\cos(\delta_{k,m})\Big/ 2s^0_{k,m}; \\ s^3_{k,m} = 2\left|E^x_{k,m}\right|\left|E^y_{k,m}\right|\sin(\delta_{k,m})\Big/ 2s^0_{k,m}. \end{cases} \quad (5)$$

В качестве иллюстрации обсуждаемого подхода рассмотрим распределения $s^0_{k,m}$–$s^3_{k,m}$ в приосевой области дальней зоны дифракции ($K_{sc} = 0.01$) для символьной последовательности, соответствующей штамму «Ухань» вируса SARS-CoV-2 [21] (рис. 1).

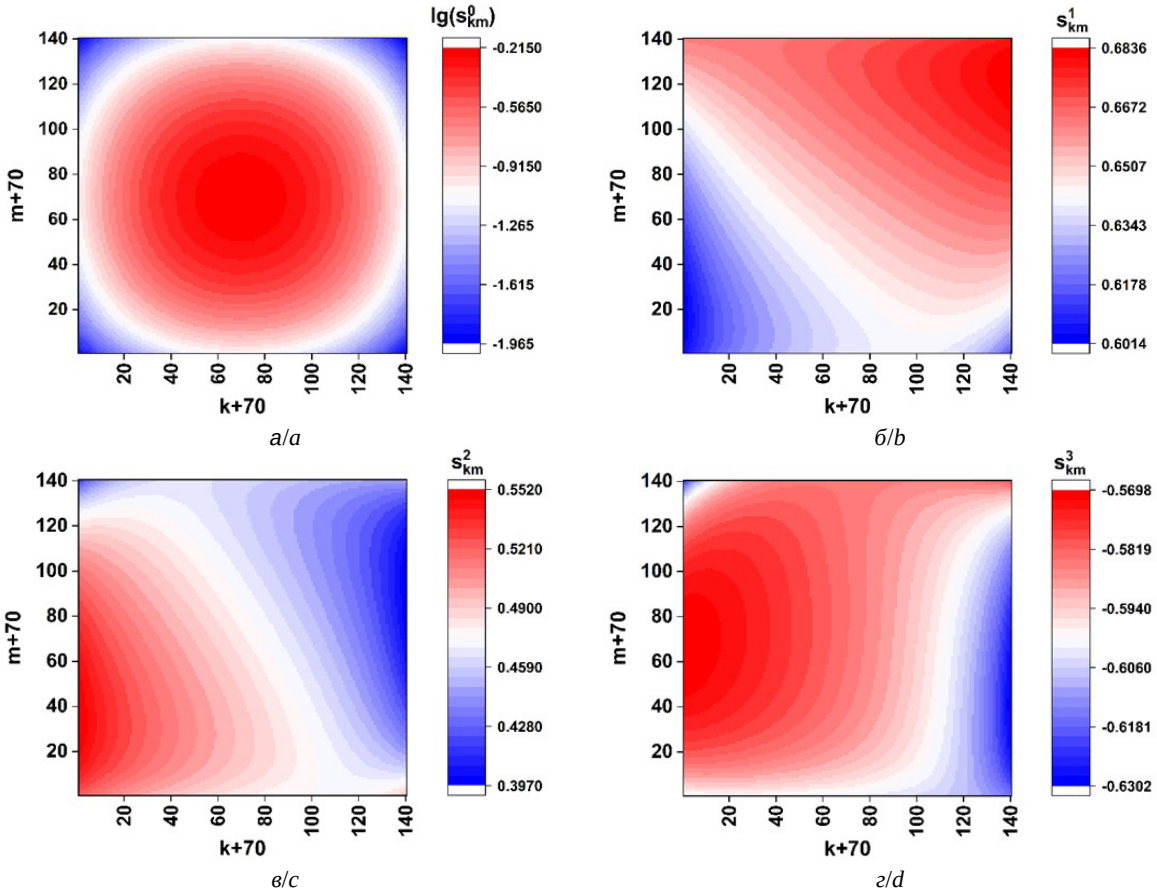
Отметим, что распределение нормированного первого компонента вектора Стокса $(s^0_{k,m})$, характеризующего полную интенсивность дифрагировавшего света (см. рис. 1, *а*) представлено в полулогарифмических координатах вследствие значительного динамического диапазона изменения данного параметра в приосевой области. При использованном значении масштабного фактора размер анализируемой области сопоставим с диаметром главного дифракционного максимума (диска Эйри).

Применительно к идентификации анализируемых символьных последовательностей рассмотрим следующий алгоритм:

$$\begin{cases} \left(\left(s^1_{k,m} < s^1_{th} + \Delta s^1_{th}\right)\ \&\ \left(s^1_{k,m} > s^1_{th} - \Delta s^1_{th}\right)\right); \\ \left(\left(s^2_{k,m} < s^2_{th} + \Delta s^2_{th}\right)\ \&\right. \\ \qquad \left.\&\ \left(s^2_{k,m} > s^2_{th} - \Delta s^2_{th}\right)\right) \rightarrow \tilde{s}_{k,m} = 1; \\ else \rightarrow \tilde{s}_{k,m} = 0. \end{cases} \quad (6)$$

В данном случае выделяются те точки анализируемой области, для которых значения нормированных второго и третьего компонентов вектора Стокса находятся внутри интервалов $s^1_{th} \pm \Delta s^1_{th}$ и $s^2_{th} \pm \Delta s^2_{th}$. Форма, площадь и положение на плоскости (k,m) зоны, выделяемой в соответствии с алгоритмом (6), уникальным образом определяются структурой фазомодулирующей матрицы $(a_{ik})_{2N\times 2N}$ и, соответственно, распределением нуклеотидов в анализируемой последовательности. Как показано ниже, незначительные изменения в распределении нуклеотидов в результате мутационных изменений приводят к изменениям этих характеристик (формы, площади и положения) зоны $(\tilde{s}_{k,m})$ для анализируемой последовательности нуклеотидов по отношению к референтной.

Выбор второго и третьего компонентов вектора Стокса как идентификационных параметров обусловлен простотой инструментальной реализации процедуры их считывания в плоскости регистрации (фокальной плоскости фурье-преобразующей линзы). Считывание значений $s^1_{k,m}$ и $s^2_{k,m}$ может быть произведено с помощью поляризатора, размещаемого перед плоскостью регистрации. Осуществляется регистрация набора значений интенсивности в точках (k,m) при



а/a *б/b* *в/c* *г/d*

Рис. 1. Цветовые карты модельных локальных состояний поляризации в приосевой (малоугловой) зоне дальней зоны дифракции для символьной последовательности, соответствующей штамму «Ухань» вируса SARS-CoV-2. Масштабный фактор K_{sc} равен 0.1: *а* – $\lg\left(s^0_{k,m}\right)$; *б* – $s^1_{k,m}$; *в* – $s^2_{k,m}$; *г* – $s^3_{k,m}$ (цвет онлайн)

Fig. 1. Color maps of model local states of polarization in the paraxial (low-angle) zone of the far diffraction zone for the symbol sequence corresponding to the Wuhan strain of the SARS-CoV-2 virus. The scale factor K_{sc} is 0.1: *a* – $\lg\left(s^0_{k,m}\right)$; *b* – $s^1_{k,m}$; *c* – $s^2_{k,m}$; *d* – $s^3_{k,m}$ (color online)

последовательных поворотах оси пропускания поляризатора на углы 0°, 45°, 90° и 135° относительно направления, соответствующего координатной оси i фазового экрана. По полученному набору значений интенсивности затем восстанавливаются значения $s^1_{k,m}$ и $s^2_{k,m}$. На рис. 2 представлено распределение $\tilde{s}_{k,m}$ для штамма «Ухань» вируса SARS-CoV-2, полученное в результате дискриминации (6) распределений $(s^1_{k,m})$ и $(s^2_{k,m})$, представленных на рис. 1, *б*, *в*.

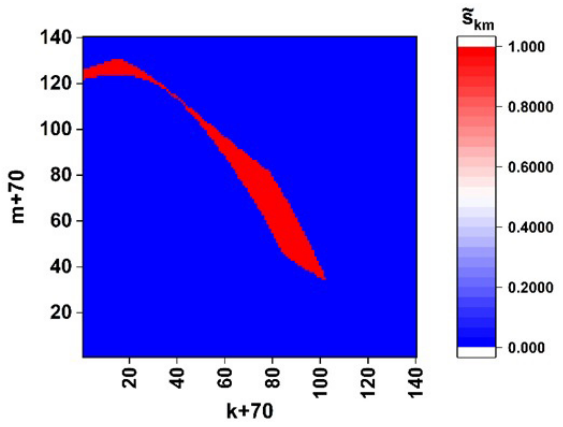


Рис. 2. Бинарное распределение приосевых локальных состояний поляризации, дискриминированных в соответствии с выражением (6) для символьной последовательности, соответствующей штамму «Ухань» вируса SARS-CoV-2. Параметры процедуры дискриминации: $s^1_{th} = 0.655$, $s^2_{th} = 0.475$, $\Delta s^1_{th} = 0.005$, $\Delta s^2_{th} = 0.005$ (цвет онлайн)

Fig. 2. Binary distribution of paraxial local polarization states discriminated according to expression (6) for the symbol sequence corresponding to the Wuhan strain of the SARS-CoV-2 virus. Discrimination procedure parameters: $s^1_{th} = 0.655$, $s^2_{th} = 0.475$, $\Delta s^1_{th} = 0.005$, $\Delta s^2_{th} = 0.005$ (color online)

Приведенные в подписи к рисунку параметры процедуры дискриминации s^1_{th}, Δs^1_{th}, s^2_{th}, Δs^2_{th} выбраны таким образом, чтобы обеспечить положение выделяемой зоны в центральной части плоскости (k,m), характеризуемой макси-

мальными значениями нормированного первого компонента вектора Стокса ($s^0_{k,m}$) (см. рис. 1).

2. Идентификация ДНК-ассоциированных символьных последовательностей с использованием дискриминированных данных малоугловой поляриметрии

Бинарные распределения $\tilde{s}_{k,m}$, получаемые в результате дискриминации (6) наборов данных малоугловой поляриметрии, могут быть применены для идентификации ДНК-ассоциированных символьных последовательности и количественного выявления различий между последовательностями, соответствующими различным штаммам одного и того же микроорганизма. Для решения данной задачи может быть применен коэффициент корреляции бинарных распределений, введенный следующим образом: $R_{1,2} = \frac{\sum_{k,m} \tilde{s}_{k,m}\tilde{s}^2_{k,m}}{\sum_{k,m}(\tilde{s}_{k,m})^2}$. Верхние индексы 1 и 2 соответствуют сравниваемым штаммам при использовании штамма «1» в качестве референтного. С учетом бинарного характера сравниваемых распределений выражение для коэффициента корреляции преобразуется к следующей форме: $R_{1,2} = \frac{\sum_{k,m} \tilde{s}_{k,m}\tilde{s}^2_{k,m}}{\sum_{k,m} \tilde{s}_{k,m}}$. Замещение части символов в анализируемой последовательности по отношению к референтной последовательности, обусловленное мутационными изменениями микроорганизма, приводит к уменьшению $R_{1,2}$. С целью оценки введенного подобным образом коэффициента корреляции к структурным изменениям последовательностей нуклеотидов (и, соответственно, к изменениям релевантных символьных последовательностей) было проведено численное моделирование поведения $R_{1,2}$ в зависимости от числа замещений в анализируемой последовательности «2» по отношению к референтной последовательности «1». В качестве «1» использована символьная последовательность для штамма «Ухань» вируса SARS-CoV-2, а анализируемая последовательность генерировалась путем случайных замещений одного, двух и более символов в «1». Для заданного числа замещений N_s, производимых случайным образом, генерировался набор случайных значений $R_{1,2}$, по которому затем вычислялось усредненное по ансамблю значение $\langle R_{1,2}\rangle$. На рис. 3 представлена полученная в результате моделирования зависимость $\langle R_{1,2}\rangle = f(N_s)$; выборочно показанные доверительные интервалы соответствуют уровню значимости 0.9.

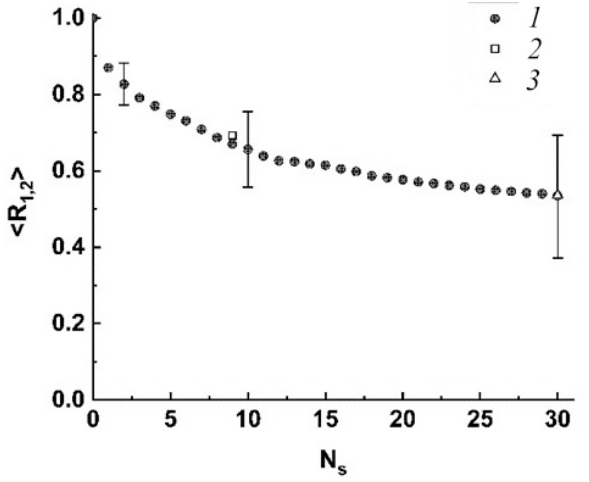


Рис. 3. Модельные значения коэффициента корреляции $\langle R_{1,2}\rangle$ в зависимости от числа замещений (*1*) и коэффициенты корреляции для пар штаммов вируса SARS-CoV-2 «Ухань – Дельта» (*2*) и «Ухань – Омикрон» (*3*) в случае малоуглового (приосевого) считывания локальных состояний поляризации и их дискриминации в соответствии с выражением (6)

Fig. 3. Model values $\langle R_{1,2}\rangle$ of the correlation coefficient depending on the number of substitutions (*1*) and correlation coefficients for pairs of strains of the SARS-CoV-2 virus “Wuhan – Delta” (*2*) and “Wuhan – Omicron” (*3*) in the case of small-angle (paraxial) reading of local polarization states and their discrimination in accordance with expression (6)

На графике также представлены значения коэффициентов корреляции бинарных распределений для пар штаммов «Ухань – Омикрон» и «Ухань – Дельта» вируса SARS-CoV-2 (символьные данные для штаммов «Омикрон» и «Дельта» заимствованы из [22, 23]). Символьная последовательность для штамма «Дельта» отличается от последовательности для штамма «Ухань» на 9 символов, а для штамма «Омикрон» на 30 символов. Соответственно, значение коэффициента корреляции для пары «Ухань – Дельта» равно $R_{1,2} \approx 0.70$, а для пары «Ухань – Омикрон» – $R_{1,2} \approx 0.54$. Отметим высокий уровень соответствия между модельными значениями коэффициента корреляции и значениями для пар «Ухань – Омикрон» и «Ухань – Дельта». Следует также отметить высокую чувствительность $R_{1,2}$ к малым изменениям в структуре символьных последовательностей (на уровне 1–2 символов).

Заключение

Таким образом, результаты проведенного моделирования позволяют сделать вывод, что регистрируемые в приосевой области дискриминированные (бинарные) распределения локаль-

ных состояний поляризации световых полей, дифрагировавших на ДНК-ассоциированных фазомодулирующих структурах, характеризуются высокой чувствительностью к локальным мутационным изменениям в структуре последовательностей нуклеотидов. Полученные результаты могут быть использованы в качестве основы при создании эффективных гибридных методов анализа генетической информации с использованием принципов поляризационного кодирования и малоугловой поляриметрии.

Следует отметить, что идея использования состояний поляризации электромагнитного излучения для передачи и обработки информации далеко не нова. В частности, начиная с восьмидесятых годов прошлого века, поляризационное кодирование и последующее декодирование бинарных последовательностей данных является одним из базовых принципов квантовой криптографии (Ч. Беннетт и Ж. Брассар, алгоритм BB84 [24], см. также [25]; Ч. Беннетт, алгоритм B92 [26] и др). В то же время обсуждаемый в данной работе подход кардинально отличается от подобных алгоритмов квантовой криптографии по следующим основным признакам:

– ассоциируемые с последовательностями нуклеотидов фазомодулирующие матрицы, равно как и формируемые в фурье-плоскости распределения локальных состояний поляризации дифрагировавшего поля, представляют собой двумерные структуры, в то время как объектом применения квантовой криптографии являются одномерные бинарные последовательности;
– число возможных состояний кодируемых элементов в триплетах равно 4, в то время как алгоритмы квантовой криптографии обрабатывают последовательности битов;
– неотъемлемой составляющей рассматриваемого подхода является двумерное фурье-преобразование синтезированных фазомодулирующих структур, приводящее к формированию в фурье-плоскости непрерывных распределений локальных состояний поляризации; лишь на заключительной стадии анализа в результате дискриминации формируются бинарные карты как идентификаторы анализируемых последовательностей нуклеотидов.

Следует отметить, что функциональность обсуждаемого подхода не ограничивается возможностью количественной идентификации структурных различий в последовательностях нуклеотидов. При модификации алгоритма фазового кодирования исходных символьных последовательностей метод малоугловой поляриметрии может быть также применен, например для частотного анализа последовательностей нуклеотидов, идентификации положений определенных нуклеотидов в последовательностях и др.

Список литературы

1. *Andelfinger G., Hitte C., Etter L., Guyon R., Bourque G., Tesler G., Pevzner P., Kirkness E., Galibert F., Benson D. W.* Detailed four-way comparative mapping and gene order analysis of the canine *ctvm* locus reveals evolutionary chromosome rearrangements // Genomics. 2004. Vol. 83. P. 1053–1062. https://doi.org/10.1016/j.ygeno.2003.12.009
2. *Anisimova M., Bielawski J. P., Yang Z.* Accuracy and power of Bayes prediction of amino acid sites under positive selection // Mol. Biol. Evol. 2002. Vol. 19. P. 950–958. https://doi.org/10.1093/oxfordjournals.molbev.a004152
3. *Rivas E., Eddy S. R.* Noncoding RNA gene detection using comparative sequence analysis // BMC Bioinform. 2001. Vol. 2. P. 1–19. https://doi.org/10.1186/1471-2105-2-8
4. *Hwang D. G., Green P.* Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution // Proc. Natl. Acad. Sci. U.S.A. 2004. Vol. 101. P. 13994–14001. https://doi.org/10.1073/pnas.0404142101
5. *Eddy S. R.* A model of the statistical power of comparative genome sequence analysis // PLoS Biol. 2005. Vol. 3. P. e10. https://doi.org/10.1371/journal.pbio.0030010
6. *Gitter A., Siegfried Z., Klutstein M., Fornés O., Oliva B., Simon I., Bar-Joseph Z.* Backup in gene regulatory networks explains differences between binding and knockout results // Mol. Syst. Biol. 2009. Vol. 5. P. 276. https://doi.org/10.1038/msb.2009.33
7. *Cooper G. M., Brudno M., Green E. D., Batzoglou S., Sidow A.* Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes // Genome Res. 2003. Vol. 13. P. 813–820. https://doi.org/10.1101/gr.1064503
8. *Abnizova I., Walter K.* Te Boekhorst R., Elgar G., Gilks W. R. Statistical information characterization of conserved non-coding elements in vertebrates // J. Bioinform. Comput. Biol. 2007. Vol. 5. P. 533–547. https://doi.org/10.1142/S0219720007002898
9. *Orlov Y. L.* Te Boekhorst R., Abnizova I. I. Statistical measures of the structure of genomic sequences: Entropy, complexity, and position information // J. Bioinform. Comput. Biol. 2006. Vol. 4. P. 523–536. https://doi.org/10.1142/S0219720006001801
10. *Sorek R., Safer H. M.* A novel algorithm for computational identification of contaminated EST libraries // Nucleic Acids Res. 2003. Vol. 31, iss. 3. P. 1067–1074. https://doi.org/10.1093/nar/gkg170

11. *Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J.* Basic local alignment search tool // J. Mol. Biol. 1990. Vol. 215. P. 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
12. Guide to Human Genome Computing / ed. M. J. Bishop. 2nd ed. San Diego, CA, USA : Academic Press, 1998. 306 p.
13. Automated DNA Sequencing and Analysis / eds. M. D. Adams, C. Fields, J. C. Venter. 1st ed. San Diego, CA, USA : Academic Press, 1994. 368 p.
14. Bioinformatics for DNA Sequence Analysis / ed. D. Posada. 1st ed. Totova, NJ, USA : Humana Press Inc., 2009. 368 p. https://doi.org/10.1007/978-1-59745-251-9
15. Оптическая голография: в 2 т. / под ред. Г. Колфилда. М. : Мир, 1982. Т. 2. 186 с.
16. *Ulianova O. V., Zaytsev S. S., Saltykov Y. V., Lyapina A., Subbotina I., Filonova N., Ulyanov S. S., Feodorova V. A.* Speckle-interferometry and speckle-correlometry of GB-speckles // Front. Biosci. (Landmark Ed). 2019. Vol. 24. P. 700–711. https://doi.org/10.2741/4744
17. *Ulyanov S. S., Ulianova O. V., Zaytsev S. S., Saltykov Y. V., Feodorova V. A.* Statistics on gene-based laser speckles with a small number of scatterers: Implications for the detection of polymorphism in the Chlamydia trachomatis omp1 gene // Las. Phys. Lett. 2018. Vol. 15, № 4. Article number 045601. https://doi.org/10.1088/1612-202X/aaa11c
18. *Goodman J. W.* Introduction to Fourier Optics. 4th ed. New York, USA : Macmillan Learning, 2017. 564 p.
19. *Goodman J. W.* Statistical Optics. 2nd ed. Hoboken, NJ, USA : J. Wiley and Sons, Inc., 2015. 544 p.
20. *Chipman R., Lam W.-S. T., Young G.* Polarized Light and Optical Systems. 1st ed. Boca-Raton, FL, USA : CRC Press, 2018. 1036 p. (Optical Sciences and Applications of Light).
21. GISAID: Official hCoV-19 Reference Sequence. URL: https://gisaid.org/wiv04/. Acc. ID: EPI_ISL_402124 (дата обращения: 15.08.2021).
22. GISAID: Official hCoV-19 Reference Sequence. URL: https://gisaid.org/wiv04/. Acc. ID: EPI_ISL_2552101 (дата обращения: 15.08.2021).
23. GISAID: Official hCoV-19 Reference Sequence. URL: https://gisaid.org/wiv04/. Acc. ID: EPI_ISL_9991311 (дата обращения: 15.08.2021).
24. *Bennett C. H., Brassard G.* Quantum cryptography: Public key distribution and coin tossing // Proceedings of International Conference on Computers, Systems & Signal Processing, Dec. 9–12, 1984, Bangalore, India. IEEE, 1984. P. 175–179.
25. *Bennett C. H., Brassard G.* Quantum cryptography: Public key distribution and coin tossing // Theoretical Computer Science. 2014. Vol. 560 (part 1). P. 7–11. https://doi.org/10.1016/j.tcs.2014.05.025
26. *Bennett C. H.* Quantum cryptography using any two nonorthogonal states // Phys. Rev. Lett. 1992. Vol. 68. P. 3121–3124. https://doi.org/10.1103/PhysRevLett.68.3121

References

1. Andelfinger G., Hitte C., Etter L., Guyon R., Bourque G., Tesler G., Pevzner P., Kirkness E., Galibert F., Benson D. W. Detailed four-way comparative mapping and gene order analysis of the canine *ctvm* locus reveals evolutionary chromosome rearrangements. *Genomics*, 2004, vol. 83, pp. 1053–1062. https://doi.org/10.1016/j.ygeno.2003.12.009
2. Anisimova M., Bielawski J. P., Yang Z. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.*, 2002, vol. 19, pp. 950–958. https://doi.org/10.1093/oxfordjournals.molbev.a004152
3. Rivas E., Eddy S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinform*, 2001, vol. 2, pp. 1–19. https://doi.org/10.1186/1471-2105-2-8
4. Hwang D. G., Green P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 2004, vol. 101, pp. 13994–14001. https://doi.org/10.1073/pnas.0404142101
5. Eddy S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.*, 2005, vol. 3, pp. e10. https://doi.org/10.1371/journal.pbio.0030010
6. Gitter A., Siegfried Z., Klutstein M., Fornés O., Oliva B., Simon I., Bar-Joseph Z. Backup in gene regulatory networks explains differences between binding and knockout results. *Mol. Syst. Biol.*, 2009, vol. 5, pp. 276. https://doi.org/10.1038/msb.2009.33
7. Cooper G. M., Brudno M., Green E. D., Batzoglou S., Sidow A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.*, 2003, vol. 13, pp. 813–820. https://doi.org/10.1101/gr.1064503
8. Abnizova I., Walter K., Te Boekhorst R., Elgar G., Gilks W. R. Statistical information characterization of conserved non-coding elements in vertebrates. *J. Bioinform. Comput. Biol.*, 2007, vol. 5, pp. 533–547. https://doi.org/10.1142/S0219720007002898
9. Orlov Y. L., Te Boekhorst R., Abnizova I. I. Statistical measures of the structure of genomic sequences: Entropy, complexity, and position information. *J. Bioinform. Comput. Biol.*, 2006, vol. 4, pp. 523–536. https://doi.org/10.1142/S0219720006001801
10. Sorek R., Safer H. M. A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.*, 2003, vol. 31, iss. 3, pp. 1067–1074. https://doi.org/10.1093/nar/gkg170
11. Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J. Basic local alignment search tool. *J. Mol. Biol.*, 1990, vol. 215, pp. 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
12. Bishop M. J., ed. *Guide to Human Genome Computing*. 2nd ed. Academic Press, San Diego, CA, USA, 1998. 306 p.
13. Adams M. D., Fields C., Venter J. C., eds. *Automated DNA Sequencing and Analysis*. 1st ed. Academic Press, San Diego, CA, USA, 1994. 368 p.

14. Posada D., ed. *Bioinformatics for DNA Sequence Analysis*. 1st ed. Humana Press Inc., Totova, NJ, USA, 2009. 368 p. https://doi.org/10.1007/978-1-59745-251-9
15. *Opticheskaya gologrefiya. Pod red. G. Colfilda* [Colfild G., ed. Optical holography: in 2 vols.]. Moscow, Mir Publ., 1982. Vol. 2. 186 p. (in Russian).
16. Ulianova O. V., Zaytsev S. S., Saltykov Y. V., Lyapina A., Subbotina I., Filonova N., Ulyanov S. S., Feodorova V. A. Speckle-interferometry and speckle-correlometry of GB-speckles. *Front. Biosci. (Landmark Ed)*, 2019, vol. 24, pp. 700–711. https://doi.org/10.2741/4744
17. Ulyanov S. S., Ulianova O. V., Zaytsev S. S., Saltykov Y. V., Feodorova V. A. Statistics on gene-based laser speckles with a small number of scatterers: Implications for the detection of polymorphism in the Chlamydia trachomatis omp1 gene. *Las. Phys. Lett.*, 2018, vol. 15, no. 4, article no. 045601. https://doi.org/10.1088/1612-202X/aaa11c
18. Goodman J. W. *Introduction to Fourier Optics*. 4th ed. Macmillan Learning, New York, USA, 2017. 564 p.
19. Goodman J. W. *Statistical Optics*. 2nd ed. J. Wiley and Sons, Inc., Hoboken, NJ, USA, 2015. 544 p.
20. Chipman R., Lam W.-S. T., Young G. *Polarized Light and Optical Systems*. 1st ed. Optical Sciences and Applications of Light. CRC Press, Boca-Raton, FL, USA, 2018, 1036 p.
21. GISAID: Official hCoV-19 Reference Sequence. Available at: https://gisaid.org/wiv04/.Acc.ID:EPI_ISL_402124 (accessed 15 Augist 2021).
22. GISAID: Official hCoV-19 Reference Sequence. Available at: https://gisaid.org/wiv04/.Acc.ID:EPI_ISL_2552101 (accessed 15 Augist 2021).
23. GISAID: Official hCoV-19 Reference Sequence. Available at: https://gisaid.org/wiv04/.Acc.ID:EPI_ISL_9991311 (accessed 15 Augist 2021).
24. Bennett C. H., Brassard G. Quantum cryptography: Public key distribution and coin tossing. *Proceedings of International Conference on Computers, Systems & Signal Processing, Dec. 9–12, 1984, Bangalore, India*. IEEE, 1984, pp. 175–179.
25. Bennett C. H., Brassard G. Quantum cryptography: Public key distribution and coin tossing. *Theoretical Computer Science*, 2014, vol. 560 (part 1), pp. 7–11. https://doi.org/10.1016/j.tcs.2014.05.025
26. Bennett C. H. Quantum cryptography using any two nonorthogonal states. *Phys. Rev. Lett.*, 1992, vol. 68, pp. 3121-3124. https://doi.org/10.1103/PhysRevLett.68.3121

Поступила в редакцию 30.10.2022; одобрена после рецензирования 02.12.2022; принята к публикации 14.12.2022
The article was submitted 30.10.2022; approved after reviewing 02.12.2022; accepted for publication 14.12.2022