Новое в прикладной физике

Известия высших учебных заведений. Прикладная нелинейная динамика. 2024. Т. 32, № 4
Izvestiya Vysshikh Uchebnykh Zavedeniy. Applied Nonlinear Dynamics. 2024;32(4)

Научная статья
УДК 004.93:577.113.5:535-4

DOI: 10.18500/0869-6632-003110
EDN: CGGWGX

Поляризационные и CGR бинарные отображения как идентификаторы нуклеотидных последовательностей в биоинформатике

Д. А. Зимняков[1,2,3]✉, *М. В. Алонова*[1], *Ан. В. Скрипаль*[2], *М. Г. Инкин*[2], *С. С. Зайцев*[4], *В. А. Федорова*[4]

[1]Саратовский государственный технический университет имени Гагарина Ю. А., Россия
[2]Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, Россия
[3]Институт проблем точной механики и управления Российской академии наук, Саратов, Россия
[4]Саратовский государственный университет генетики, биотехнологии и инженерии имени Н. И. Вавилова, Россия
E-mail: ✉zimnykov@mail.ru, alonova_marina@mail.ru, skripalav@info.sgu.ru, sunbeam18.95@mail.ru, zaytsev-sergey@inbox.ru, feodorovav@mail.ru

Поступила в редакцию 7.11.2023, принята к публикации 28.02.2024, опубликована онлайн 28.05.2024, опубликована 31.07.2024

Аннотация. *Цель* данного исследования — сравнительный анализ двух подходов к синтезу двумерных бинарных идентификаторов нуклеотидных последовательностей, получаемых в результате секвенирования ДНК различных биологических объектов. *Методы*. Один из подходов основан на моделировании процесса поляризационно-зависимой дифракции считывающего когерентного пучка на двумерной фазомодулирующей структуре (фазовом экране), ассоциируемой с получаемой в результате секвенирования ДНК символьной последовательностью. В основе другого подхода используется двумерное представление анализируемой последовательности с использованием популярного отображения хаотической игры (chaos game representation, CGR). Для получения конечно-размерного CGR отображения производится его фрагментация на заданное число ячеек, обеспечивающее приемлемую чувствительность синтезированного бинарного идентификатора к структурным изменениям отображаемой последовательности. *Результаты*. Сравнительный анализ проведен с использованием символьных последовательностей, соответствующих различным штаммам («Ухань», «Дельта», «Омикрон») вируса SarSCoV2, для которых получены и сравнены значения коэффициентов взаимной корреляции синтезированных бинарных идентификаторов. *Заключение*. Установлено, что бинарные идентификаторы, синтезированные с использованием метода поляризационного кодирования, характеризуются существенно более высокой чувствительностью к структурным изменениям анализируемых последовательностей и меньшими размерами в сравнении с CGR идентификаторами.

Ключевые слова: нуклеотидные последовательности, бинарное отображение, поляризационное кодирование, представление хаотической игры.

Благодарности. Работа выполнена при поддержке РНФ, грант № 22-21-00194.

Для цитирования: *Зимняков Д. А., Алонова М. В., Скрипаль Ан. В., Инкин М. Г., Зайцев С. С., Федорова В. А.* Поляризационные и CGR бинарные отображения как идентификаторы нуклеотидных последовательностей в биоинформатике // Известия вузов. ПНД. 2024. Т. 32, № 4. С. 439–459. DOI: 10.18500/0869-6632-003110

Article DOI: 10.18500/0869-6632-003110

Polarization- and CGR-based binary representations as identifiers of the nucleotide sequences in bioinformatics

D. A. Zimnyakov[1,2,3]✉, *M. V. Alonova*[1], *An. V. Skripal*[2], *M. G. Inkin*[2], *S. S. Zaitsev*[4], *V. A. Feodorova*[4]

[1]Yury Gagarin State Technical University of Saratov, Russia
[2]Saratov State University, Russia
[3]Institute for Problems of Precision Mechanics and Control of the Russian Academy of Sciences, Saratov, Russia
[4]Saratov State University of Genetics, Biotechnology and Engineering named after N. I. Vavilov, Russia
E-mail: ✉zimnykov@mail.ru, alonova_marina@mail.ru, skripalav@info.sgu.ru, sunbeam18.95@mail.ru, zaytsev-sergey@inbox.ru, feodorovav@mail.ru
Received 7.11.2023, accepted 28.02.2024, available online 28.05.2024, published 31.07.2024

Abstract. *Purpose* of this work is the comparative analysis of two approaches to the synthesis of two-dimensional binary identifiers of nucleotide sequences obtained using DNA sequencing of biological objects. *Methods*. One of the approaches is based on modeling the polarization-dependent diffraction of a coherent readout beam on a two-dimensional phase-modulating structure (phase screen) associated with the symbolic sequence obtained as a result of DNA sequencing. Another approach uses a two-dimensional representation of the symbolic sequence using a chaos game representation (CGR). To obtain a finite-element CGR mapping, it is fragmented into a given number of cells, ensuring acceptable sensitivity of the synthesized binary identifier to structural changes in the displayed sequence. *Results*. The comparative analysis was carried out using fragments of symbol sequences corresponding to various strains (Wuhan, Delta, Omicron) of the SarSCoV2 virus. In the course of the analysis, the correlation coefficients between the binary identifiers corresponding to various strains were obtained and compared with each other. *Conclusion*. It has been established that binary identifiers synthesized using the polarization encoding technique are characterized by significantly higher sensitivity to structural changes in the analyzed sequences and smaller sizes compared to CGR binary identifiers.

Keywords: nucleotide sequences, binary representation, polarization encoding, chaos game representation.

Acknowledgements. This work was supported by the Russian Science Foundation, grant No. 22-21-00194.

For citation: Zimnyakov DA, Alonova MV, Skripal AnV, Inkin MG, Zaitsev SS, Feodorova VA. Polarization- and CGR-based binary representations as identifiers of the nucleotide sequences in bioinformatics. Izvestiya VUZ. Applied Nonlinear Dynamics. 2024;32(4):439–459. DOI: 10.18500/0869-6632-003110

Введение

Анализ и визуализация генетической информации, получаемой в результате секвенирования фрагментов ДНК и РНК различных биологических объектов [1], является одним из доминирующих направлений фундаментальной и прикладной исследовательской деятельности в современной биоинформатике. Объектами исследований в данном случае являются символьные последовательности различной длины (от нескольких сотен до нескольких сотен тысяч символов) и их фрагменты, характеризуемые квазислучайными распределениями четырех символов (A, C, T, G). Символы A, C, T, G ассоциируются с базовыми нуклеотидами, формирующими структуру анализируемого фрагмента ДНК (A — аденин, C — цитозин, T — тимин, G — гуанин [2]), а порядок их расположения в последовательности является уникальным атрибутом данного биологического объекта. Квазислучайность распределений нуклеотидов в секвенируемых фрагментах ДНК (и, соответственно, ассоциированных символов в последовательностях) обусловлена как существованием крупномасштабных корреляций между положениями нуклеотидов в ДНК, так и различиями в относительных частотах нахождения базовых нуклеотидов при случайной выборке их позиций в последовательностях.

За прошедшие три десятилетия в биоинформатике сформировалась общая стратегия синтеза отображений ДНК-ассоциированных символьных последовательностей в двумерном или трехмерном декартовом пространстве. Эта стратегия основана на выборе четырехточечного базиса, каждая точка которого определенным образом соотносится с одним из четырех символов. В настоящее время в основном используются три типа базисов, различающихся порядком расположения символов при обходе точек базиса по часовой стрелке: RY-базис (ACGT), MK-базис (AGCT) и WS-базис (ACTG). Выбор того или иного базиса производится исходя из возможностей получения дополнительной информации о биохимических особенностях анализируемого фрагмента ДНК; в частности, отображение последовательности в RY-базисе позволяет осуществить частотный анализ нуклеотидов по пуриновым и пиримидиновым группам. Аденин и гуанин (A, G; побочная диагональ базиса) относятся к классу пиримидинов, в то время как цитозин и тимин (C, T; главная диагональ базиса) принадлежат к классу пуринов. Аналогичным образом отображение в MK-базисе позволяет проанализировать распределение нуклеотидов по аминовым и кетоновым группам, а WS-базис предоставляет возможность в части анализа распределения нуклеотидов по группам со слабыми и сильными водородными связями [3–5].

Последовательное построение образов последовательностей производится с помощью различных рекурсивных процедур; в результате синтезируются либо кусочно-линейные отображения в трехмерном пространстве (например, Z-, C-, RY-, MK- и WS-линии [6–10]), либо точечные отображения в двумерном пространстве. В качестве широко используемого двумерного точечного отображения нуклеотидных последовательностей следует отметить отображение хаотической игры (chaos game representation, CGR), впервые предложенное Джеффри (H. Joel Jeffrey, [11]). В соответствии с CGR-алгоритмом синтез двумерного точечного отображения символьной последовательности осуществляется в пределах квадратной области единичных размеров, где символу A соответствует начало координат (0.0; 0.0). Остальным точкам базиса соответствуют координаты (0.0; 1.0), (1.0; 1.0) и (1.0; 0.0). Соотнесение оставшихся символов (C, G, T) и этих точек производится исходя из выбранного базиса (RY, MK или WS). В качестве начальной точки выбирается центр квадрата (0.5; 0.5) и синтез отображения производится в соответствии с расположением символов в последовательности по правилу, что очередная отображающая точка располагается на середине отрезка, соединяющего предшествующую точку и точку базиса, соответствующую отображаемому символу. Эта рекурсивная процедура продолжается вплоть до окончания символьной последовательности. Существуют определенные аналогии между данным алгоритмом и алгоритмами синтеза двумерных фрактальных структур (например, ковра Серпинского [12]). Существование отмеченных выше крупномасштабных корреляций в положениях символов в последовательностях при достаточно больших их длинах приводит к фракталоподобным структурам синтезируемых CGR отображений некоторых биологических объектов.

Существенным недостатком бинарных CGR-карт при малых длинах символьных последовательностей (от нескольких сотен до нескольких тысяч символов) является малая средняя поверхностная плотность отображающих точек, не позволяющая надежно идентифицировать, например, фрактальный характер их структуры. В связи с этим в последние два десятилетия были предложены различные модификации данного подхода, в том числе и FCGR-алгоритм (frequency chaos game representation, [13–16]). Основная идея данной модификации заключается в разделении CGR отображения на равновеликие ячейки, подсчете числа отображающих точек в пределах каждой ячейки и характеризации ячеек по относительным частотам попадания в них точек. Соответственно, относительная частота как информативный параметр может отображаться в градациях серого цвета [17].

Очевидно, что изменения в структуре отображаемой символьной последовательности по отношению к референтной последовательности, обусловленные мутационными замещениями

нуклеотидов в анализируемом фрагменте ДНК, приводят к смещениям части отображающих точек в синтезированной CGR-карте по отношению к опорной (референтной) карте. Другими словами, синтезированная карта может рассматриваться как уникальный идентификатор символьной последовательности и, соответственно, данного фрагмента ДНК. Уровень мутационных изменений может быть количественно оценен путем оценки степени взаимной корреляции анализируемой и опорной бинарных CGR структур.

В работах [17–19] был предложен альтернативный подход к синтезу двумерных бинарных отображений нуклеотидных последовательностей на основе моделирования эффекта поляризационной модуляции когерентного светового пучка ДНК-ассоциированным двумерным фазовым экраном. Подобная модуляция приводит к формированию в дальней зоне дифракции распределения локальных состояний поляризации дифрагировавшего пучка, описываемых тремя компонентами нормированного вектора Стокса. Было показано [18], что двумерные бинарные отображения, формируемые путем дискриминации пространственных распределений локальных значений компонентов вектора Стокса, также могут рассматриваться в качестве идентификаторов нуклеотидных последовательностей, обладающих высокой чувствительностью к мутационным изменениям.

Целью данной работы является сравнительный анализ двух подходов к двумерной бинарной идентификации фрагментов нуклеотидных последовательностей, один из которых основан на методике синтеза CGR-карт, а второй использует принцип поляризационного кодирования нуклеотидных последовательностей и выявления предельных состояний поляризации дифрагировавшего когерентного светового поля. В качестве анализируемых объектов рассмотрены фрагменты символьных последовательностей, ассоциированных с тремя различными штаммами вируса SARS-CoV-2 (ковида) — «Ухань», «Дельта» и «Омикрон». Вирус SARS-CoV-2 продолжает представлять высокую потенциальную угрозу для человечества [20] вследствие своей высокой антигенной изменчивости [21]. Среди генов, кодирующих белки SARS-CoV-2, значительный интерес для исследований представляет спайк-ликопротеин, или S-белок, оказывающий влияние на проникновение вируса в клетки хозяина [22, 23]. Более того, S-белок может распространяться в организме и отдельно от вируса, высвобождаясь из инфицированных вирусных частиц, определяясь в разных органах и тканях, повреждая клетки макроорганизма; фрагменты этого белка могут проникать сквозь гематоэнцефалический барьер [24]. Следует также отметить высокую частоту мутаций нуклеотидной последовательности кодирующего биосинтез S-белка [25] по сравнению с более консервативными генами, экспрессирующими синтез других полипептидов SARS-CoV-2. Нуклеотидная последовательность гена «S», кодирующего спайк-белок, расположена в геноме вируса в позициях 21563–25384 и имеет длину 3822 нуклеотидов (1274 аминокислоты) [26]. Выбор штаммов «Ухань», «Дельта» и «Омикрон» вируса SARS-CoV-2 обусловлен тем, что штамм «Ухань» обычно рассматривается в качестве референтного, а штаммы «Дельта» и «Омикрон», равно как и их сублинии, рассматриваются как представляющие большую угрозу в сравнении с другими штаммами в силу их «наибольшей заразности» [27].

1. Описание анализируемых символьных последовательностей

Символьные последовательности взяты из находящейся в открытом доступе базы данных GISAID (Global Initiative on Sharing All Influenza Data, открытый доступ по подписке), в которой штамму «Ухань» соответствует ссылка [28], «Дельта» — [29], и «Омикрон» — [30]. Число A, C, T, G символов во фрагментах последовательностей, ассоциируемых с геном «S», равно 3822.

Соответственно, число триплетов (аминокислот) в анализируемых фрагментах равно 1274. Табл. 1 отображает различия в последовательностях триплетов между тремя рассматриваемыми штаммами; штамм «Ухань» обычно рассматривается в качестве референтного; различающиеся триплеты отмечены символом (!).

Таким образом, символьная последовательность для штамма «Дельта» отличается от референтной последовательности для штамма «Ухань» 9 триплетами, в то время как в последовательности для штамма «Омикрон» имеются 30 различающихся триплетов. Отметим, что все различия обусловлены мутационными замещениями единичных нуклеотидов в каждом из различающихся триплетов.

Таблица 1. Различия в последовательностях триплетов для трех отображаемых штаммов

Table 1. Differences in the triplet sequences for three displayed strains

Положение триплетов в последовательностях	«Ухань»	«Дельта»	«Омикрон»
19	ACA	AGA(!)	ATA(!)
95	ACT	ATT(!)	ACT
142	GGT	GAT(!)	GAT(!)
213	GTG	GTG	GGG(!)
339	GGT	GGT	GAT(!)
371	TCC	TCC	TTC(!)
373	TCA	TCA	CCA(!)
375	TCC	TCC	TTC(!)
376	ACT	ACT	GCT(!)
405	GAT	GAT	AAT(!)
408	AGA	AGA	AGC(!)
410	ATC	ATC	ATT(!)
417	AAG	AAT(!)	AAT(!)
440	AAT	AAT	AAG(!)
452	CTG	CGG(!)	CTG
477	AGC	AGC	AAC(!)
478	ACA	AAA(!)	AAA(!)
484	GAA	GAA	GCA(!)
493	CAA	CAA	CGA(!)
498	CAA	CAA	CGA(!)
501	AAT	AAT	TAT(!)
505	TAC	TAC	CAC(!)
614	GAT	GGT(!)	GGT(!)
655	CAT	CAT	TAT(!)
679	AAT	AAT	AAG(!)
681	CCT	CGT(!)	CAT(!)
764	AAC	AAC	AAA(!)
796	GAT	GAT	TAT(!)
925	AAC	AAC	AAT(!)
950	GAT	AAT(!)	GAT
954	CAA	CAA	CAT(!)
969	AAT	AAT	AAA(!)
1146	GAC	GAC	GAT(!)

2. Поляризационное кодирование ДНК-ассоциированных символьных последовательностей и синтез двумерных бинарных идентификаторов

Как отмечено выше, методика виртуального поляризационного кодирования ДНК-ассоциированных символьных последовательностей, рассмотренная в [18], заключается в представлении анализируемой последовательности или ее фрагмента двумерной фазомодулирующей структурой (фазовым экраном, содержащим $2\tilde{N}_t \times 2\tilde{N}_t$ элементов, где $\tilde{N}_t$ — количество триплетов в анализируемой структуре). Фазовый экран считывается коллимированным когерентным пучком с линейным состоянием поляризации, и в дальней зоне дифракции (фокальной плоскости фурье-преобразующей линзы) анализируются пространственные распределения локальных состояний поляризации дифрагировавшего пучка. На рис. 1 представлена физическая интерпретация моделируемой процедуры считывания ДНК-ассоциированного фазового экрана.

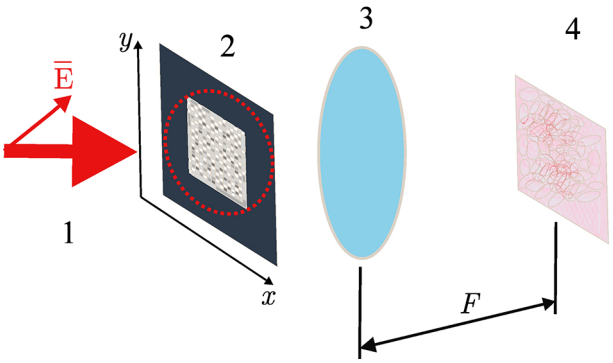


Рис. 1. Физическая интерпретация процедуры поляризационного кодирования ДНК-ассоциированной символьной последовательности и синтеза двумерного бинарного идентификатора. ДНК-ассоциированный фазовый экран (2) освещается линейно-поляризованным коллимированным лазерным пучком (1) с плоскостью поляризации, ориентированной под углом 45° к сторонам экрана (пунктирной линией показан освещаемый участок экрана). Поляризационно-зависимая дифракционная структура (4) формируется в фокальной плоскости фурье-преобразующей линзы. На рисунке не показан блок поляризационного анализа, расположенный между линзой (3) и фурье-плоскостью (4) и используемый для формирования пространственных распределений $(s^0_{k,m} \div s^3_{k,m})$

Fig. 1. Physical interpretation of the procedure for polarization encoding of a DNA-associated symbol sequence and the synthesis of a two-dimensional binary identifier. The DNA-associated phase screen (2) is illuminated by a linearly polarized collimated laser beam (1) with a plane of polarization oriented at an angle of 45° to the sides of the screen (the dotted line marks the illuminated area of the screen). The polarization-dependent diffraction structure (4) is formed in the focal plane of the Fourier-transforming lens. The figure does not show the polarization-analyzing unit, located between the lens (3) and the Fourier plane (4) and used to form spatial distributions $(s^0_{k,m} \div s^3_{k,m})$

Локальные состояния поляризации определяются значениями компонентов вектора Стокса $(s^0_{k,m} \div s^3_{k,m})$. При синтезе бинарного идентификатора кодированной символьной последовательности определяются пространственные положения в фурье-плоскости предельных состояний четвертого компонента вектора Стокса $s^3_{k,m}$, характеризующего вклад циркулярно-поляризованной составляющей в состояние поляризации дифрагировавшего пучка в точке (k, m) фурье-плоскости. Отметим, что в рамках используемого формализма дискретному набору $2\tilde{N}_t \times 2\tilde{N}_t$ локальных фазовых сдвигов, вносимых синтезированным фазовым экраном в считывающий когерентный пучок, соответствуют $4\tilde{N}_t \times 4\tilde{N}_t$ точек (пикселей) фурье-плоскости, однозначно определяемых двумерным дискретным преобразованием Фурье. В качестве критерия выбора предельных состояний $s^3_{k,m}$ используется условие $|s^3_{\text{th}}| \leqslant |s^3_{k,m}| < |\pm 1|$, где пороговое значение s^3_{th} выбирается близким к -1 (в случае дискриминации левоциркулярных состояний поляризации), либо к 1 (при дискриминации правоциркулярных состояний). Выбор $s^3_{k,m}$ в качестве идентификационного параметра обусловлен тем, что близкие к циркулярному состояния поляризации характеризуются максимальной чувствительностью к локальным изменениям структуры синтезированного фазового экрана по сравнению с линейно-поляризованными составляющими дифрагировавшего поля, определяемыми компонентами $s^1_{k,m}$ и $s^2_{k,m}$ [18].

При синтезе ДНК-ассоциированного фазового экрана он генерируется как ансамбль $\tilde{N}_t \times \tilde{N}_t$ субматриц размером (2×2), каждая из которых ассоциируется с определенным триплетом в отображаемой последовательности A, C, T, G символов. Очевидно, что размер $\tilde{N}_t \times \tilde{N}_t$ отображаемого набора триплетов в последовательности определяется максимальным значением квадрата целого числа, не превосходящим числа триплетов в последовательности. Для иллюстрации данного положения рассмотрим символьную последовательность, соответствующую «S»-гену штамма «Ухань» вируса SARS-CoV-2. В секвенированном участке ДНК присутствует 3822 нуклеотидов (соответственно, 1274 триплета, см. раздел 1). Соответственно, максимально возможное значение $\tilde{N}_t$ при синтезе ДНК-ассоциированного фазового экрана равно 35. $\tilde{N}_t^2 = 1225$ и 49 триплетов не включены в синтез фазового экрана; однако, следуя данным, представленным в табл. 1, можно заключить, что отбрасываемые фрагменты последовательностей для штаммов «Ухань», «Дельта» и «Омикрон» не содержат каких-либо различий.

Каждая субматрица в $\tilde{N}_t \times \tilde{N}_t$ наборе отображает свойства набора базовых нуклеотидов в соответствующем триплете в соответствии с выбранным правилом кодирования. В качестве примера возможной взаимосвязи между элементами субматриц $(b_{0,0} \div b_{2,2})$ и базовыми нуклеотидами рассмотрим следующее правило:

$$b_{0,0} \to A;\ b_{1,0} \to C;\ b_{0,1} \to T;\ b_{1,1} \to G. \tag{1}$$

Значение элемента определяется числом нуклеотидов данного типа в триплете; соответственно, универсальное правило для всех субматриц в наборе заключается в том, что сумма их элементов всегда равна 3. В качестве иллюстрации рис. 2 отображает в градациях серого структуру синтезированного фазового экрана для «S»-гена штамма «Ухань», используемого в качестве референтного объекта.

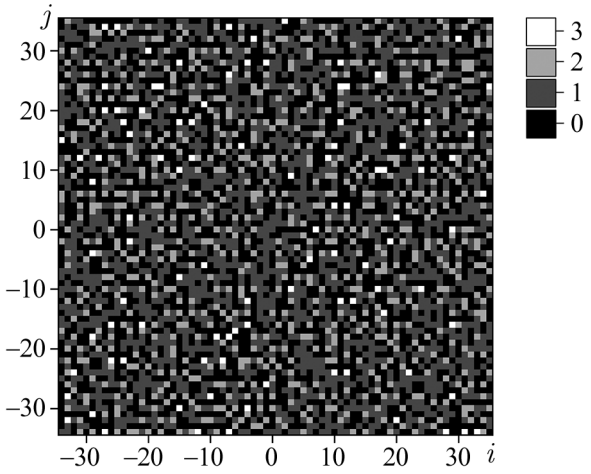


Рис. 2. Структура ДНК-ассоциированного фазового экрана для «S»-гена штамма «Ухань»

Fig. 2. Structure of the DNA-associated phase screen for the “S” gene of the “Wuhan” strain

Матрица фазового экрана $(a_{i,j})_{2\tilde{N}_t \times 2\tilde{N}_t}$, сформированная путем последовательной построчной и постолбцовой комбинации субматриц $(b_{i',j'})_{2\times 2}$, соответствующих триплетам, используется для фазовой модуляции считывающего линейно-поляризованного пучка. Поскольку в считывающем пучке присутствуют x- и y-поляризованные составляющие с равными значениями амплитуды и начальной фазы (плоскость поляризации считывающего пучка образует углы 45° со сторонами фазового экрана), то, применяя закон модуляции фазы следующего вида:

$$\begin{aligned} (\Delta\varphi_{ij})^x_{2\tilde{N}_t \times 2\tilde{N}_t} &= \pi(a_{i,j})_{2\tilde{N}_t \times 2\tilde{N}_t}, \\ (\Delta\varphi_{ij})^y_{2\tilde{N}_t \times 2\tilde{N}_t} &= 0.5\pi + \pi(a_{i,j})_{2\tilde{N}_t \times 2\tilde{N}_t}, \end{aligned} \tag{2}$$

получим за непосредственно экраном набор из $2\tilde{N}_t \times 2\tilde{N}_t$ локальных состояний циркулярной поляризации граничного светового поля, различия между которыми обусловлены двумя возможными случайными значениями фазы x-поляризованной составляющей — 0 (когда $a_{i,j} = 0$ или 2) и π (когда $a_{i,j} = 1$ или 3). «Перемешивание» прошедших через различные элементы фазового экрана парциальных составляющих считывающего когерентного поля в дальней зоне дифракции

может быть описано двумерным дискретным преобразованием Фурье x- и y-поляризованных составляющих граничного поля (см., например, [31]):

$$E^{x,y}_{k,m} = \frac{1}{4\tilde{N}_t^2} \sum_{i=-\tilde{N}_t}^{\tilde{N}_t-1} \sum_{j=-\tilde{N}_t}^{\tilde{N}_t-1} \exp[-\tilde{j} \cdot K_{sc} \cdot \{(\pi/\tilde{N}_t)(k \cdot i + m \cdot j) - \Delta\varphi^{x,y}_{ij}\}], \tag{3}$$

где $\tilde{j}$ — мнимая единица, индексы k, m определяют дискретные координаты пикселей в фурье-плоскости, а коэффициент K_{sc} определяет масштаб отображения пространственного распределения дифрагировавшего считывающего пучка в фурье-плоскости. Максимальное допустимое значение K_{sc} равно 0.5 и соответствует наиболее крупномасштабному («панорамному») отображению фурье-спектра фазового экрана. Использование значений масштабного коэффициента K_{sc}, превышающих 0.5, приводит к искажению моделируемых пространственных распределений амплитуды x- и y-поляризованных составляющих дифракционного поля вследствие эффекта подмены частот [32]. Уменьшение K_{sc} соответствует большей детализации отображения центральной области дифрагировавшего поля с одновременным уменьшением размеров отображаемого участка. Отметим, что при записи выражения (3) было принято допущение о единичных значениях амплитуд x- и y-поляризованных составляющих считывающего пучка, не нарушающее общности проводимого рассмотрения.

Для полученных подобным образом модельных дискретных распределений $E^{x,y}_{k,m}$ затем производится вычисление компонентов вектора Стокса в соответствии со следующими выражениями (см., например, [33]):

$$\begin{cases} s^0_{k,m} = (|E^x_{k,m}|^2 + |E^y_{k,m}|^2)/2, \\ s^1_{k,m} = (|E^x_{k,m}|^2 - |E^y_{k,m}|^2)/2s^0_{k,m}, \\ s^2_{k,m} = 2|E^x_{k,m}||E^y_{k,m}|\cos(\delta_{k,m})/2s^0_{k,m}, \\ s^3_{k,m} = 2|E^x_{k,m}||E^y_{k,m}|\sin(\delta_{k,m})/2s^0_{k,m}, \end{cases} \tag{4}$$

где $\delta_{k,m}$ — значения разностей фаз x- и y-поляризованных составляющих в соответствующих точках фурье-плоскости. Нормированные значения $(s^1_{k,m} \div s^3_{k,m})$ принимают значения в интервалах от -1 до 1 и удовлетворяют фундаментальному соотношению поляризационной оптики:

$$(s^1_{k,m})^2 + (s^2_{k,m})^2 + (s^3_{k,m})^2 = 1. \tag{5}$$

Таким образом, при значительном вкладе правой или левой циркулярно-поляризованных составляющих в состояние поляризации дифрагировавшего пучка в точке k, m фурье-плоскости имеют место следующие соотношения между нормированными компонентами вектора Стокса: $s^3_{k,m} \to \pm 1;\ s^1_{k,m}, s^2_{k,m} \to 0$.

В качестве примера на рис. 3 представлены отображения распределений значений $s^3_{k,m}$ для «S»-ассоциированных символьных последовательностей, соответствующих штаммам «Ухань», «Дельта» и «Омикрон»; при их построении использовано значение масштабного коэффициента K_{sc}, равное 0.1 (детализированное отображение). Для идентификации различий между распределениями $s^3_{k,m}$, обусловленными замещениями части нуклеотидов (и, соответственно, символов) в анализируемых последовательностях, они могут быть бинаризованы в соответствии со следующими правилами:

$$\begin{cases} 1 \geqslant s^3_{k,m} \geqslant s^3_{\text{th}} \to \tilde{s}^3_{k,m} = 1, \\ s^3_{k,m} < s^3_{\text{th}} \to \tilde{s}^3_{k,m} = 0, \end{cases} \tag{6}$$

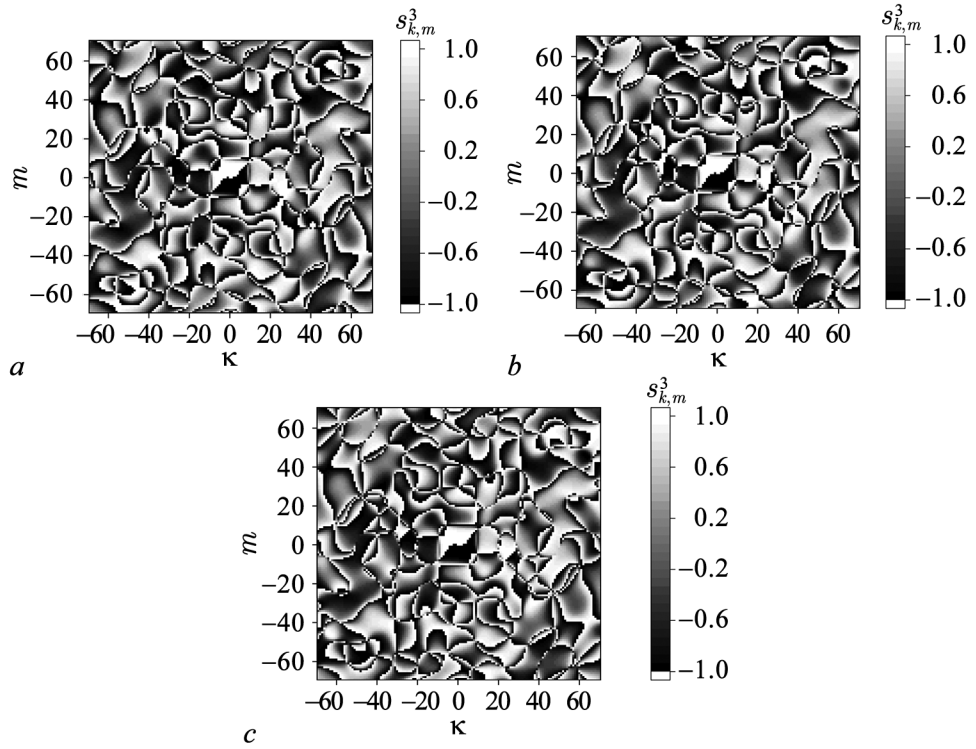


Рис. 3. Детализированные ($K_{sc} = 0.1$) отображения нормированных значений четвертого компонента вектора Стокса для штаммов «Ухань» (*a*), «Дельта» (*b*) и «Омикрон» (*c*)

Fig. 3. Detailed (small-scale) ($K_{sc} = 0.1$) representations of normalized values of the fourth component of the Stokes vector for the “Wuhan” (*a*), “Delta” (*b*) and “Omicron” (*c*) strains

в случае выделения предельных состояний, близких к правой циркулярной поляризации, и, соответственно,

$$\begin{cases} -1 \leqslant s^3_{k,m} \leqslant s^3_{\text{th}} \to \tilde{s}^3_{k,m} = 1, \\ s^3_{k,m} > s^3_{\text{th}} \to \tilde{s}^3_{k,m} = 0, \end{cases} \tag{7}$$

при выделении предельных состояний, близких к левой циркулярной поляризации. Пороговое значение s^3_{th} выбирается близким к 1 (для правой циркулярной поляризации) или к -1 (соответственно, для левой циркулярной поляризации).

На рис. 4 представлены результаты применения подобной процедуры к распределениям, представленным на рис. 3 с порогом дискриминации, равным -0.98. Сопоставление бинарных распределений может быть произведено путем попиксельного логического перемножения референтного (штамм «Ухань») и анализируемого распределения. Результаты этой процедуры представлены на рис. 5. Количественно степень соответствия анализируемого и референтного идентификаторов может быть выражена с помощью коэффициента корреляции

$$R^{a,r} = \frac{\sum\limits_{m=1}^{4\tilde{N}_t} \sum\limits_{k=1}^{4\tilde{N}_t} a_{m,k} \times r_{m,k}}{\sum\limits_{m=1}^{4\tilde{N}_t} \sum\limits_{k=1}^{4\tilde{N}_t} r_{m,k}}, \tag{8}$$

где символы $a_{m,k}$ и $r_{m,k}$ относятся соответственно к пикселям анализируемого и референтного идентификаторов.

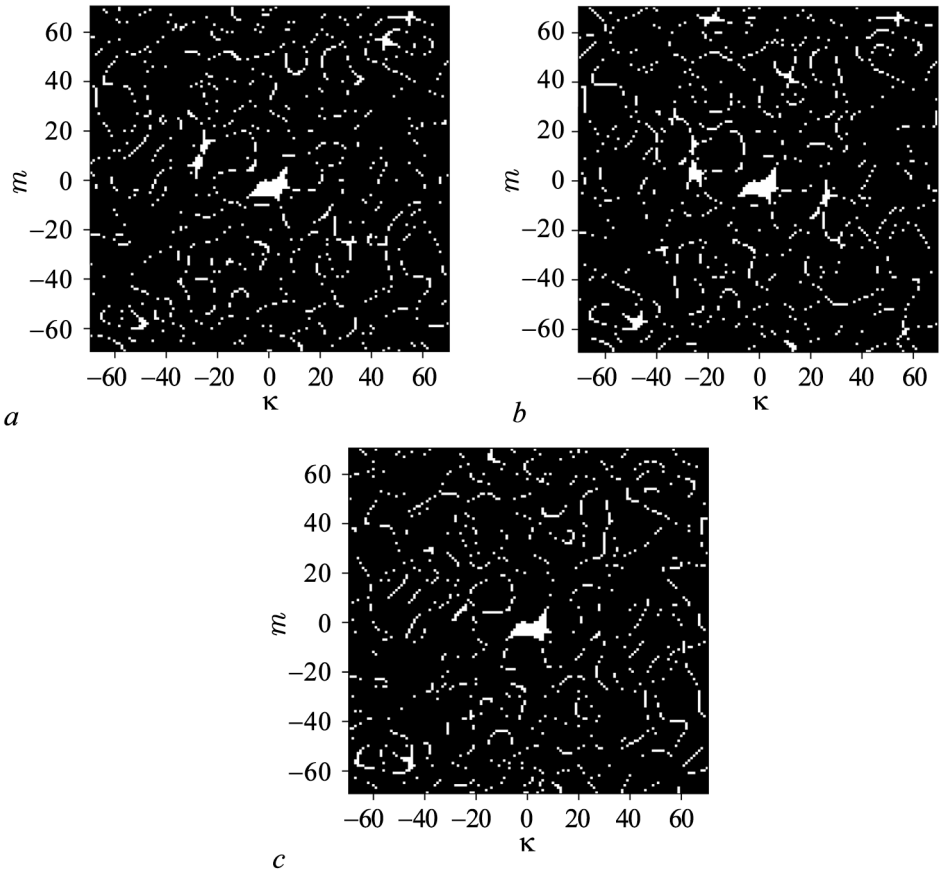


Рис. 4. Бинарные идентификаторы «S»-генов штаммов «Ухань» (*a*), «Дельта» (*b*) и «Омикрон» (*c*), полученные в результате поляризационного кодирования. Выделены локальные состояния, близкие к левой циркулярной поляризации, с порогом дискриминации $s_{\text{th}}^3 = -0.98$

Fig. 4. Binary identifiers of “S” genes of strains “Wuhan” (*a*), “Delta” (*b*) and “Omicron” (*c*), obtained using polarization encoding. Local polarization states close to the left circular polarization are shown for the discrimination threshold $s_{\text{th}}^3 = -0.98$

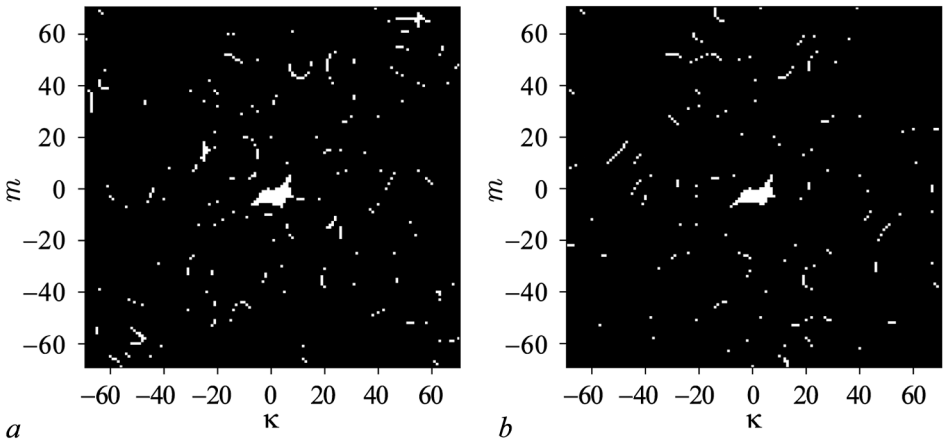


Рис. 5. Результаты попиксельного логического перемножения бинарного идентификатора для референтной последовательности (штамм «Ухань», рис. 4, *a*) и бинарных идентификаторов для штаммов «Дельта» (*a*) и «Омикрон» (*b*)

Fig. 5. Results of pixel-by-pixel logical multiplication of a binary identifier for the reference sequence (“Wuhan” strain, Fig. 4, *a*) and binary identifiers for the “Delta” (*a*) and “Omicron” (*b*) strains

3. CGR отображения ДНК-ассоциированных символьных последовательностей

CGR-карты как двумерные дискретные отображения символьных последовательностей, рассматриваемые в качестве альтернативы описанному выше методу поляризационного бинарного кодирования, синтезируются в квадратной области с единичной стороной в соответствии со следующей схемой. В зависимости от используемого базиса (RY, MK или WS), вершины квадрата с декартовыми координатами (0, 0), (0, 1), (1, 1) и (1, 0) (направление обхода — по часовой стрелке) определенным образом ассоциируются с четырьмя базовыми нуклеотидами. Для RY-базиса имеют место следующие ассоциации: A → (0,0), C → (0,1), G → (1,1), T → (1,0); для MK-базиса — A → (0,0), G → (0,1), C → (1,1), T → (1,0), и для WS-базиса — A → (0,0), C → (0,1), T → (1,1), G → (1,1). Исходная точка для построения выбирается в центре квадрата ((0.5, 0.5)). Первый символ в последовательности отображается точкой, соответствующей середине отрезка между исходной точкой и вершиной квадрата, соответствующей отображаемому символу.

Следующий символ отображается точкой, соответствующей середине отрезка между предшествующей отображающей точкой и вершиной квадрата, ассоциируемой с символом. Эта рекурсивная процедура повторяется вплоть до завершающего символа последовательности. В качестве примера подобной процедуры рассмотрим результат генерации координат отображающих точек для первых 5 членов «S»-ассоциированной символьной последовательности для штамма «Ухань» (ATGTT. . . .) в RY-базисе: (0.5;0.5), (0.25;0.25), (0.625;0.125), (0.813;0.563), (0.906;0.281), (0.953;0.141). Формально алгоритм генерации отображающих точек может быть представлен следующими выражениями:

$$\begin{cases} x_n = \frac{1}{2^{n+1}} + \sum\limits_{k=1}^{n} \dfrac{C_x^k}{2^{n-k+1}}, \\ y_n = \frac{1}{2^{n+1}} + \sum\limits_{k=1}^{n} \dfrac{C_y^k}{2^{n-k+1}}, \end{cases} \tag{9}$$

где коэффициенты C_x^k и C_y^k определяют координаты вершины базиса, соответствующей k-му нуклеотиду в последовательности. Очевидно, что отображающие точки всегда находятся внутри единичного квадрата или по крайней мере на его границах.

На рис. 6 в качестве примера представлена CGR-карта символьной последовательности для «S»-гена штамма «Ухань» в RY-базисе. Следует отметить существенную неоднородность пространственного распределения отображающих точек (наличие как зон с высокой локальной плотностью заполнения, так и лакун). Для некоторых нуклеотидных последовательностей подобная пространственная неоднородность позволяет говорить о «фрактальности» CGR отображения, в определенном смысле аналогичной, например, классическому ковру Серпинского. В качестве примера можно сослаться на работу [34], наглядно иллюстрирующую фрактальный характер CGR отображений большой длины (порядка нескольких сотен тысяч нуклеотидов).

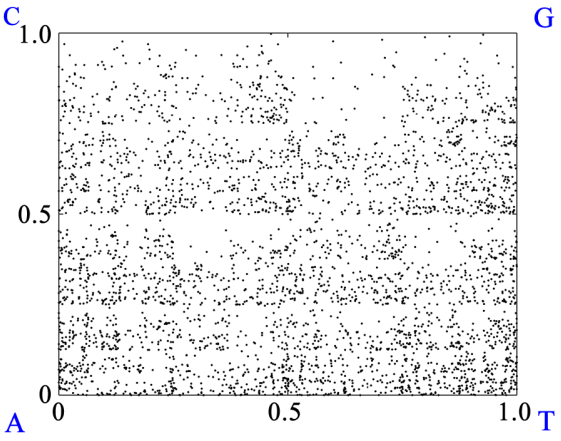


Рис. 6. CGR отображение символьной последовательности для «S»-гена штамма «Ухань» в RY-базисе

Fig. 6. CGR map of the symbolic sequence for the “S” gene of the “Wuhan” strain in the RY basis

Учитывая квазинепрерывный характер распределений x- и y-координат отображающих точек в синтезируемых CGR-отображениях символьных наборов, анализ

корреляций отображений референтной и анализируемой последовательностей в данном случае отличается от рассмотренного выше алгоритма логического перемножения бинарных матриц относительно малого размера, как в случае поляризационного кодирования (выражение (8)). Следует отметить, что квазинепрерывный характер распределений координат отображающих точек в случае CGR-отображений не согласуется с концепцией синтеза двумерных бинарных идентификаторов относительно небольших размеров для ДНК-ассоциируемых символьных последовательностей. Решение данной проблемы сводится к фрагментации («огрублению») CGR-карт, что в определенной степени аналогично упомянутому выше FCGR-алгоритму, хотя и существенно отличается от него. На качественном уровне это решение формулируется следующим образом: необходимо определить минимально допустимый уровень разбиения CGR-карты на равновеликие ячейки, при котором вероятность попадания двух и более отображающих точек в произвольно выбранный фрагмент меньше заданного порогового значения. Другими словами, внутри каждого фрагмента может либо находиться единственная отображающая точка, либо нет.

На рис. 7 представлены результаты статистического анализа процесса «огрубления» CGR-карт для штамма «Ухань» в RY-, MK- и WS-базисах в зависимости от параметра N/N_n (N — число разбиений координатных осей, соответственно, число фрагментов CGR-карты равно N^2; N_n — число символов в отображаемой последовательности). Кривые *1–3* отображают значения вероятности нахождения нескольких (больше чем 1) отображающих точек P_s в ячейках RY, MK и WS CGR-карт, выделяемых путем случайной выборки, в зависимости от N/N_n. Для сравнения также приведена аналогичная зависимость (4) для случая однородного распределения отображающих точек по CGR-карте. Отметим значительно большие значения W для анализируемой символьной последовательности в сравнении с равномерным распределением, обусловленные существенно неоднородными пространственными распределениями отображающих точек в синтезированных RY, MK и WS CGR-картах (см. рис. 6 для RY-отображения).

На рис. 8 представлены рассчитанные значения коэффициента корреляции $R^{a,r}$ (см. выражение (8)) между референтной (штамм «Ухань») и анализируемыми (штаммы «Дельта» (*1*)

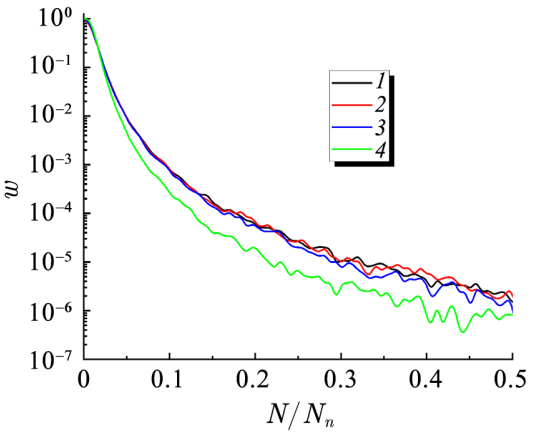


Рис. 7. Модельные значения вероятности обнаружения ячейки CGR-карты, содержащей более одной отображающей точки, в зависимости от уровня фрагментации карты N/N_n. *1* — RY-базис; *2* — MK-базис; *3* — WS-базис; *4* — однородное распределение отображающих точек по площади карты (цвет онлайн)

Fig. 7. Model values of the probability of detecting a CGR map cell containing more than one mapping point, depending on the level of map fragmentation N/N_n. *1* — RY basis; *2* — MK basis; *3* — WS basis; *4* — uniform distribution of mapping points over the map area (color online)

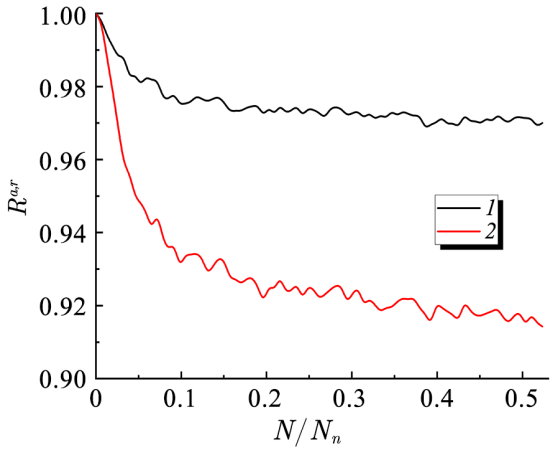


Рис. 8. Коэффициенты корреляции между бинарными идентификаторами «S»-ассоциированных референтной (штамм «Ухань») и анализируемых последовательностей (штамм «Дельта» (*1*); штамм «Омикрон» (*2*)) в зависимости от N/N_n (RY-базис) (цвет онлайн)

Fig. 8. Correlation coefficients between binary identifiers of “S”-associated reference (strain “Wuhan”) and analyzed sequences (strain “Delta” (*1*); strain “Omicron” (*2*)) depending on N/N_n (RY basis) (color online)

и «Омикрон» (*2*)) фрагментированными CGR-картами в зависимости от отношения N/N_n. Следует отметить, что зависимости $R^{a,r} = f(N/N_n)$ при значениях N/N_n, превышающих 0.3, демонстрируют практически постоянные значения коэффициента корреляции, что позволяет предложить данный уровень фрагментации как минимальный при синтезе бинарных идентификаторов на основе фрагментированных CGR-карт. Вероятность обнаружения ячеек с более чем одной отображающей точкой, которым при фрагментации приписываются единичные значения, невелика и составляет приблизительно 2×10^{-5}. С учетом числа символов в анализируемых последовательностях (3822), число разбиений каждой из 2 координатных осей CGR-карты составляет не менее 1147. Соответственно, общее число пикселей в синтезированном подобным образом бинарном идентификаторе составляет не менее 1315609. Отметим, что применение других базисов (MK, WS) для построения исходных CGR-карт приводит к аналогичным результатам.

4. Обсуждение результатов

Сравнительный анализ двух рассмотренных подходов применительно к синтезу двумерных бинарных идентификаторов ДНК-ассоциированных символьных последовательностей целесообразно провести путем сравнения двух ключевых параметров: число элементов (пикселей) в идентификаторе и чувствительность бинарной структуры идентификатора к структурным изменениям в анализируемой символьной последовательности по отношению к референтной последовательности. В качестве последнего параметра целесообразно рассмотреть коэффициент корреляции $R^{a,r}$, определяемый выражением (8). В случае поляризационного кодирования символьных последовательностей число пикселей в синтезируемых идентификаторах определяется числом используемых триплетов $\tilde{N}_t^2$ и равно $16\tilde{N}_t^2$. В то же время коэффициент корреляции $R^{a,r}$ существенным образом зависит от используемого порога дискриминации s_{th}^3 предельных состояний поляризации.

В случае же синтеза бинарных идентификаторов на основе фрагментированных CGR-карт как число пикселей, так и коэффициенты корреляции между референтным и анализируемыми идентификаторами зависят от используемого уровня фрагментации. Эти особенности иллюстрируются данными, представленными в табл. 2 и 3.

Сопоставление приведенных наборов данных позволяет сделать вывод о существенных преимуществах бинарных идентификаторов, синтезированных с использованием поляризационного кодирования, по сравнению с идентификаторами, получаемых путем фрагментации CGR отображений. Возрастание по абсолютной величине порога дискриминации в первом случае приводит не только к повышению чувствительности коэффициента корреляции к замещениям

Таблица 2. Параметры бинарных идентификаторов, синтезированных на основе поляризационного кодирования

Table 2. Parameters of binary identifiers synthesized on the base of polarization encoding

Порог дискриминации s_{th}^3	Число пикселей в бинарном идентификаторе	Коэффициент корреляции $R^{a,r}$ для пары «Ухань – Дельта»	Коэффициент корреляции $R^{a,r}$ для пары «Ухань – Омикрон»
-0.92	19600	≈ 0.49	≈ 0.41
-0.94	19600	≈ 0.45	≈ 0.36
-0.96	19600	≈ 0.41	≈ 0.31
-0.98	19600	≈ 0.34	≈ 0.25
-0.999	19600	≈ 0.21	≈ 0.13

Таблица 3. Параметры бинарных идентификаторов, синтезированных на основе фрагментации CGR-карт

Table 3. Parameters of binary identifiers synthesized on the base of fragmentation of the CGR maps

Уровень фрагментации N/N_n	Число пикселей в бинарном идентификаторе	Коэффициент корреляции $R^{a,r}$ для пары «Ухань – Дельта»	Коэффициент корреляции $R^{a,r}$ для пары «Ухань – Омикрон»
0.05	36864	≈ 0.982	≈ 0.951
0.1	146689	≈ 0.977	≈ 0.935
0.2	585225	≈ 0.973	≈ 0.926
0.3	1315609	≈ 0.972	≈ 0.926
0.4	2337841	≈ 0.972	≈ 0.919
0.5	3651921	≈ 0.972	≈ 0.918

отдельных нуклеотидов в анализируемой последовательности по отношению к референтной, но также и к уменьшению числа пикселей с единичными значениями в идентификаторе. Так, при пороге дискриминации, равном -0.92, число пикселей с единичными значениями равно 2256, в то время как для порога дискриминации, равного -0.999, это число равно 266. Ранее [18] отмечалось, что метод поляризационного кодирования следует применять для сравнения нуклеотидных последовательностей, различающихся в относительно небольших количествах нуклеотидов (от 1 до 10). В частности, при искусственных случайных замещениях одного из символов в рассматриваемом фрагменте последовательности для штамма «Ухань» (3822 символа) и пороге дискриминации, равном -0.99, значение коэффициента корреляции $R^{a,r}$ между бинарными поляризационными идентификаторами исходного и модифицированного фрагментов равно $\approx 0.556 \pm 0.036$ [18]. Среднее значение $R^{a,r}$ и доверительный интервал, соответствующий уровню значимости 0.9, были получены для выборки из 30 случайным образом модифицированных фрагментов. Отметим, что почти двукратное уменьшение коэффициента корреляции (с 1.0 до 0.556) происходит при замене только одного символа, ассоциированного с одним из 4 базовых нуклеотидов (то есть при различии в структурах исходной и модифицированной последовательностей порядка 0.026%). В соответствии с данными табл. 2, при возрастании абсолютной величины порога дискриминации до значений, близких к 1, чувствительность $R^{a,r}$ к изменениям в структуре последовательностей существенно увеличивается. С другой стороны, рассматриваемый фрагмент последовательности, соответствующий S-белку в геноме SARS-CoV-2, характеризуется относительно малой длиной. Вопрос о предельной чувствительности бинарных поляризационных идентификаторов к малым изменениям в структуре длинных последовательностей (порядка нескольких десятков тысяч элементов и более) будет исследован в дальнейшем.

На основе разработанных алгоритмов синтеза бинарных поляризационных и CGR-идентификаторов ДНК-ассоциированных символьных последовательностей были реализованы пробные тексты программ на языке C++. Программы обеспечивали считывание исходных символьных последовательностей произвольной длины из текстового файла (*.txt), конвертацию символьных (char) данных в целочисленные (integer) значения, обработку полученных массивов целочисленных данных с одновременным преобразованием их к формату двойной точности с плавающей точкой (double) и вывод в текстовый файл (*.dat) синтезированных бинарных идентификаторов, представляющих собой двумерные массивы целочисленных (integer) данных. В соответствии с описанными выше методами синтеза бинарных идентификаторов, размер формируемой квадратной матрицы определяется длиной исходной символьной последовательности, а элементы матрицы являются однобитовыми величинами, принимая значения либо 1, либо 0. Синтез

поляризационных идентификаторов осуществлялся в 2 этапа: конвертация символьной последовательности в фазомодулирующую матрицу (1, исходная программа convert.cpp, исполняемый файл convert.exe); синтез соответствующего бинарного распределения предельных состояний циркулярной поляризации в фурье-плоскости (2, исходная программа pol_map.cpp, исполняемый файл pol_map.exe). При синтезе CGR идентификаторов использовался единый исполняемый файл cgr_map.exe (соответствующий исходный текст cgr_map.cpp). Отметим, что программы представляют собой пилотные версии, предназначенные только для верификации обсуждаемых в данной работе методов бинарного отображения нуклеотидных последовательностей. Возможна их оптимизация с целью экономии вычислительных ресурсов (объема используемой оперативной памяти и времени вычислений), однако эти вопросы выходят за пределы настоящего исследования и планируются к проведению в будущем.

Тем не менее были проведены оценки временных затрат на синтез бинарных поляризационных и CGR идентификаторов с использованием исполняемых файлов convert.exe, pol_map.exe и cgr_map.exe. В качестве исходных данных были использованы фрагменты символьных последовательностей различной длины (в интервале от 3822 до 492 символов), получаемые в результате пошагового исключения из исходной «S»-ассоциированной последовательности для штамма «Ухань» концевых групп из 333 символов (111 триплетов). Тестирование исполняемых файлов производилось на персональном компьютере с процессором AMD Ryzen 9 3900X (12 ядер, 3.79 ГГц) и 32 Гб оперативной памяти. На рис. 9 представлены значения времени выполнения файлов convert.exe, pol_map.exe, cgr_map.exe и размеры синтезируемых поляризационных и CGR идентификаторов в зависимости от длины символьной последовательности. Следует отметить, что для относительно коротких символьных последовательностей (с числом элементов до ≈ 1000) значения времени выполнения для всех программ сопоставимы и определяются главным образом длительностью процедур загрузки исходных символьных данных с жесткого диска в оперативную память и вывода синтезированных бинарных идентификаторов из оперативной памяти на жесткий диск. С увеличением числа символов в последовательностях время выполнения программы pol_map.exe существенно возрастает по сравнению с временем выполнения cgr_map.exe (приблизительно пятикратное различие для 3822 символов). Это обусловлено использованием достаточно длительной процедуры вычисления двумерного дискретного преобразования Фурье при синтезе поляризационных идентификаторов по сравнению с существенно более быстрым распределением точек CGR отображения по ячейкам бинарных CGR идентификаторов. Однако следует иметь в виду, что, несмотря на более времязатратный процесс, бинарные поляризационные идентификаторы характеризуются существенно более высокой чувствительностью к структурным изменениям анализируемых последовательностей (см. табл. 2, 3) и существенно меньшими размерами (рис. 9) по сравнению с бинарными CGR идентификаторами. Также были синтезированы бинарные поляризационные и CGR идентификаторы для полного генома штамма «Ухань», описываемо-

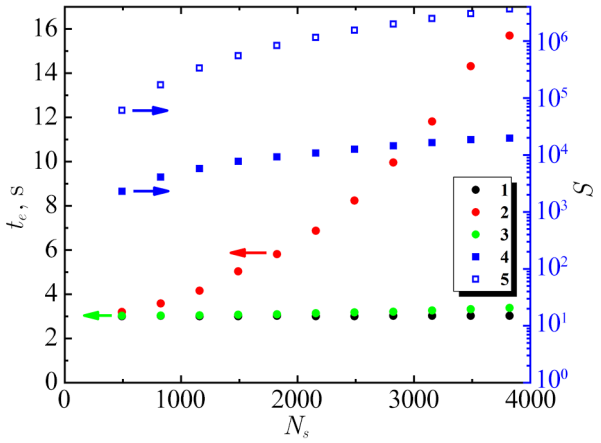


Рис. 9. Время выполнения t_e файлов convert.exe (1), pol_map.exe (2), cgr_map.exe (3) и размеры S бинарных поляризационных (4) и CGR (5) идентификаторов в зависимости от числа символов в обрабатываемой последовательности. Стрелки показывают отнесение данных к осям (цвет онлайн)

Fig. 9. The execution time t_e for the files convert.exe (1), pol_map.exe (2), cgr_map.exe (3) and the sizes S of binary polarization (4) and CGR (5) identifiers depending on the number of symbols in the processed sequence. Arrows show the assignment of data to axes (color online)

го 29891 символом [28]. В первом случае время синтеза составило 849 с при размере идентификатора 396×396 пикселей; в случае же CGR идентификатора время синтеза равно 24.85 с при размере 14945×14945 пикселей. Отметим, что в тестируемом файле pol_map.exe применена достаточно простая «линейная» процедура обработки считываемых символьных данных, которая может быть оптимизирована с целью уменьшения времени вычислений (например, в результате использования алгоритма быстрого преобразования Фурье (БПФ) вместо обычного дискретного преобразования Фурье, распараллеливания вычислительного процесса и т. д.).

Одной из ключевых проблем применения рассмотренного в работе метода синтеза бинарных поляризационных идентификаторов является частичная потеря информации о структуре последовательности при ее преобразовании в квадратную фазомодулирующую матрицу с меньшим числом элементов, чем число триплетов анализируемой последовательности. В рассмотренных примерах с использованием символьных последовательностей для штаммов «Ухань», «Дельта» и «Омикрон» вируса SARS-CoV-2 эта проблема не имеет значения, поскольку все различия в структурах штаммов проявляются в пределах выделяемого участка, содержащего 1225 триплетов (3675 символов). Однако для других объектов может существовать определенная вероятность попадания структурных несоответствий во фрагменты последовательностей, отбрасываемые при преобразовании. На наш взгляд, существуют два возможных пути обхода этой проблемы в дальнейших исследованиях, первый из которых заключается в соотнесении с анализируемой последовательностью не одного, а двух бинарных поляризационных идентификаторов одинакового размера, один из которых синтезируется при считывании фрагмента символьной последовательности, начиная с первого символа в прямом направлении, а второй — при считывании фрагмента такой же длины в обратном направлении, начиная с последнего символа. Таким образом, особенности структуры последовательности будут в полной мере отображаться этой парой бинарных идентификаторов. Другой подход может быть основан на представлении структуры последовательности в форме не квадратной, а прямоугольной фазомодулирующей матрицы, размеры которой определяются в результате представления числа триплетов в последовательности в виде произведения двух целых чисел, близких друг к другу по величине. Соответственно, синтезированный подобным образом бинарный идентификатор будет также представлять собой прямоугольную матрицу. В предельном случае, когда число триплетов в символьной последовательности является простым, возможно построение «вырожденного» поляризационного идентификатора в виде матрицы размером $4 \times 4N_T$, где N_T — полное число триплетов в последовательности. Время синтеза подобного идентификатора будет существенно меньше в сравнении с временем формирования двумерного идентификатора. Верификация и сравнительный анализ этих подходов, равно как и оптимизация алгоритмов и программ синтеза бинарных поляризационных идентификаторов, являются объектами дальнейших исследований.

Процедура синтеза бинарных поляризационных идентификаторов фрагментов нуклеотидных последовательностей может быть реализована не только с применением только компьютерной обработки данных, но и в рамках гибридного (инструментально-программного подхода). При этом, в соответствии с рис. 1, считывание ДНК-ассоциированной фазомодулирующей матрицы осуществляется коллимированным пучком линейно-поляризованного непрерывного лазерного излучения, в качестве источника которого может быть использован, например, одномодовый гелий-неоновый лазер ГН-5П. В качестве фазомодулирующего экрана 2 может быть применен компьютерно-управляемый жидкокристаллический пространственный модулятор фазы типа LS2012 с разрешением 1024×768 пикселей и 8-битовым представлением кодируемой информации (производство компании Holoeye Photonics AG, Германия). Считывание ДНК-ассоциированных дифракционных картин в фокальной плоскости фурье-преобразующей линзы 3 может осуществляться с использованием поляризационно-чувствительной КМОП камеры Kiralux CS505MUP1 (разрешение — 2448×2448 пикселей, 12-битовое представление считываемых данных,

официальный дистрибьютор — компания Thorlabs, США). Размер пикселей жидкокристаллической матрицы LS2012 равен 36 мкм; камера Kiralux CS505MUP1 имеет размер пикселей, равный 3.45 мкм. В случае дискретного преобразования Фурье структуры с размером элементов, равным Δ, соответствующий размер элементов в фурье-образе структуры Δ_r, масштабный коэффициент K_{sc}, фокусное расстояние F линзы 3 и длина волны лазерного света λ связаны между собой соотношением $K_{sc} = \Delta\Delta_r/F\lambda$. Рассмотрим случай инструментальной реализации поляриметрической системы рис. 1 с использованием жидкокристаллического модулятора LS2012, камеры Kiralux CS505MUP1 и гелий-неонового лазера ($\lambda = 0.63$ мкм). При поляризационном отображении символьных последовательностей для штаммов «Ухань», «Дельта» и «Омикрон» вируса SARS-CoV-2 (число элементов фазомодулирующей матрицы 70×70) рабочая зона фазомодулирующего экрана занимает в его центральной части размеры 2.52 мм$\times$2.52 мм. Предположим, что анализируемые поляризационно-чувствительные дифракционные паттерны, содержащие 140×140 элементов в случае представления дискретным преобразованием Фурье, отображаются на всю рабочую зону камеры размером 8.45 мм$\times$8.45 мм. При коэффициенте масштабирования, равном $K_{sc} = 0.1$, фокусное расстояние фурье-преобразующей линзы должно быть равно ≈ 34 мм. Отметим, что область фурье-плоскости, соответствующая одному элементу дискретного фурье-образа, перекрывается при этом приблизительно тремястами пикселей камеры. Апертура линзы 3 должна превышать размеры рабочей зоны фазомодулирующего экрана; соответственно, одним из возможных решений может являться использование коммерчески доступных линз, поставляемых компанией Thorlabs (например, LA1700 с диаметром 6 мм и фокусным расстоянием 30 мм) или российских аналогов. Меньшее фокусное расстояние приведет к небольшому возрастанию коэффициента масштабирования до значения 1.13, что не окажет существенного влияния на функционирование системы. Следует также отметить, что при использовании в качестве источников считывающего излучения одномодовых гелий-неоновых лазеров, традиционно применяемых в интерферометрических, дифрактометрических и поляриметрических измерениях, каким-либо влиянием амплитудно-фазовых шумов считывающего пучка на поляризационно-зависимое дифракционное отображение символьных последовательностей можно пренебречь.

Заключение

Таким образом, рассмотренный метод поляризационного кодирования демонстрирует высокую эффективность в части синтеза двумерных бинарных объектов, уникальным образом отображающих структуру ДНК-ассоциированных символьных последовательностей. Следует отметить, что возможности данного подхода не ограничиваются рассмотренным в данной работе алгоритмом преобразования структуры символьной последовательности в структуру фазомодулирующей матрицы (фазового экрана). Например, рассмотренный в разделе 2 алгоритм кодирования субматриц исходя из содержания базовых нуклеотидов в соответствующих триплетах (см. выражение (1)) при условии обхода субматриц по часовой стрелке соответствует RY-базису. Изменение ассоциаций элементов субматриц при поляризационном кодировании в соответствии с правилами: $b_{0,0} \to A$; $b_{1,0} \to G$; $b_{0,1} \to C$; $b_{1,1} \to T$ и $b_{0,0} \to A$; $b_{1,0} \to C$; $b_{0,1} \to G$; $b_{1,1} \to T$ (соответственно, привязка MK- и WS-базисов к кодированию) позволит расширить функциональные возможности рассматриваемого подхода. В частности, анализ влияния выбора базиса при поляризационном кодировании является объектом дальнейших исследований в данном направлении. Другим возможным направлением исследований является структурный анализ синтезированных бинарных отображений при значениях порога дискриминации, близких к ± 1 (аналогичных представленным на рис. 4). Например, в качестве элементов, характеризующих подобные структуры, могут быть рассмотрены точки разрывов и ветвлений линий предельных состояний поляризации, проявляющихся на бинарных отображениях рис. 4.

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;17(6):333–351. DOI:10.1038/nrg.2016.49.
2. Neidle S, Sanderson M. Principles of Nucleic Acid Structure. Academic Press; 2021. 454 p.
3. Randic M, Vracko M, Lers N, Plavsic D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters. 2003;368(1–2):1–6. DOI: 10.1016/S0009-2614(02)01784-0.
4. Randic M, Vracko M, Nandy A, Basak SC. On 3-D graphical representation of DNA primary sequence and their numerical characterization. Journal of Chemical Information and Computer Sciences. 2000;40(5):1235–1244. DOI: 10.1021/ci000034q.
5. Xie G, Mo Z. Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. Journal of Theoretical Biology. 2011;269(1):123–130. DOI: 10.1016/j.jtbi.2010.10.018.
6. Jafarzadeh N, Iranmanesh A. A novel graphical and numerical representation for analyzing DNA sequences based on codons. Match-Communications in Mathematical and Computer Chemistry. 2012;68(2):611–620.
7. Jafarzadeh N, Iranmanesh A. C-curve: A novel 3D graphical representation of DNA sequence based on codons. Mathematical Biosciences. 2013;241(2):217–224. DOI: 10.1016/j.mbs.2012.11.009.
8. Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. Journal of Biological Chemistry. 1983;258(2):1318–1327. DOI: 10.1016/S0021-9258(18)33196-X.
9. Zhang CT, Zhang R, Ou HY. The Z-curve databases: A graphic representation of genome sequence. Bioinformatics. 2003;19(5):593–599. DOI: 10.1093/bioinformatics/btg041.
10. Yu ZG, Wang B. A time series model of CDS sequences in complete genome. Chaos Solitons Fractals. 2001;12(3):519–526. DOI: 10.1016/S0960-0779(99)00208-8.
11. Jeffrey HJ. Chaos game representation of gene structure. Nucleic Acids Research. 1990;18(8):2163–2170. DOI: 10.1093/nar/18.8.2163.
12. Anitas EM. Small-angle scattering and multifractal analysis of DNA sequences. International Journal of Molecular Sciences. 2020;21(13):4651. DOI: 10.3390/ijms21134651.
13. Burma PK, Raj A, Deb JK, Brahmachari SK. Genome analysis: a new approach for visualization of sequence organization in genomes. Journal of Biosciences. 1992;17(4):395–411. DOI: 10.1007/BF02720095.
14. Huynen MA, Konings DAM, Hogeweg P. Equal G and C contents in histone genes indicate selection pressures on mRna secondary structure. Journal of Molecular Evolution. 1992;34(4):280–291. DOI: 10.1007/BF00160235.
15. Hill KA, Schisler NJ, Singh SM. Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. Journal of Molecular Evolution. 1992;35(3):261–269. DOI: 10.1007/BF00178602.
16. Almeida JS, Carrico JA, Maretzek A, Noble PA, Fletcher M. Analysis of genomic sequences by chaos game representation. Bioinformatics. 2001;17(5):429–437. DOI: 10.1093/bioinformatics/17.5.429.
17. Zimnyakov DA, Alonova MV, Skripal AnV, Zaitsev SS, Feodorova VA. Polarization analysis of gene sequence structures: Mapping of extreme local polarization states. Journal of Biomedical Photonics & Engineering. 2022;8(4):040302. DOI: 10.18287/JBPE22.08.040302.
18. Zimnyakov DA, Alonova MV, Skripal AnV, Dobdin SY, Feodorova VA. Quantification of the

diversity in gene structures using the principles of polarization mapping. Current Issues in Molecular Biology. 2023;45(2):1720–1740. DOI: 10.3390/cimb45020111.
19. Ulyanov SS, Ulianova OV, Zaytsev SS, Saltykov YV, Feodorova VA. Statistics on gene-based laser speckles with a small number of scatterers: implications for the detection of polymorphism in the Chlamydia trachomatis omp1 gene. Laser Physics Letters. 2018;15:045601. DOI: 10.1088/1612-202X/aaa11c.
20. Rak A, Isakova-Sivak I, Rudenko L. Overview of Nucleocapsid-Targeting Vaccines against COVID-19. Vaccines. 2023;11(12):1810. DOI: 10.3390/vaccines11121810.
21. Telenti A, Hodcroft EB, Robertson DL. The Evolution and Biology of SARS-CoV-2 Variants. Cold Spring Harbor Perspectives in Medicine. 2022;12:a041390. DOI: 10.1101/cshperspect.a041390.
22. Bergmann CC, Silverman RH. COVID-19: coronavirus replication, pathogenesis, and therapeutic strategies. Cleveland Clinic Journal of Medicine. 2020;87:321—327 DOI: 10.3949/ccjm.87a.20047.
23. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F. Cell entry mechanisms of SARS-CoV-2. Proceedings of the National Academy of Sciences. 2020;117:11727—11734. DOI: 10.1073/pnas.2003138117.
24. Grobbelaar LM, Venter C, Vlok M, Ngoepe M, Laubscher GJ, Lourens PJ, Steenkamp J, Kell DB, Pretorius E. SARS-CoV-2 spike protein S1 induces fibrin (ogen) resistant to fibrinolysis: implications for microclot formation in COVID-19. Bioscience Reports. 2021;41(8):BSR20210611. DOI: 10.1042/BSR20210611.
25. Singh D, Yi SV. On the origin and evolution of SARS-CoV-2. Experimental & Molecular Medicine. 2021;53:537—547. DOI: 10.1038/s12276-021-00604-z.
26. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270–273. DOI: 10.1038/s41586-020-2012-7.
27. Chakraborty C, Bhattacharya M, Chopra H, Bhattacharya P, Islam MA, Dhama K. Recently emerged omicron subvariant BF.7 and its R346T mutation in the RBD region reveal increased transmissibility and higher resistance to neutralization antibodies: need to understand more under the current scenario of rising cases in China and fears of driving a new wave of the COVID-19 pandemic. International Journal of Surgery. 2023;109(4):1037–1040. DOI: 10.1097/JS9.0000000000000219.
28. GISAID [Electronic resource]: Official hCoV-19 Reference Sequence. Acc. ID: EPI_ISL_402124. Available online: https://gisaid.org/wiv04/.
29. GISAID [Electronic resource]: Official hCoV-19 Reference Sequence. Acc. ID: EPI_ISL_2552101. Available online: https://gisaid.org/wiv04/.
30. GISAID [Electronic resource]: Official hCoV-19 Reference Sequence. Acc. ID: EPI_ISL_9991311. Available online: https://gisaid.org/wiv04/.
31. Goodman JW. Introduction to Fourier Optics, 4th ed. New York: Macmillan Learning; 2017. 491 p.
32. Bracewell R. The Fourier Transform and Its Applications. New York: McGraw Hill; 1986. 474 p.
33. Chipman R, Lam WST, Young G. Polarized Light and Optical Systems (Optical Sciences and Applications of Light). Boca-Raton: CRC Press; 2018. 1036 p.
34. Anitas EM. Fractal analysis of DNA sequences using frequency chaos game representation and small-angle scattering. International Journal of Molecular Sciences. 2022;23(3):1847. DOI: 10.3390/ijms23031847.

Зимняков Дмитрий Александрович — окончил Саратовский государственный университет им. Н. Г. Чернышевского по специальности «Оптика» (1978). Защитил диссертации на соискание ученой степени кандидата технических наук (1989) и доктора физико-математических наук (1997). Имеет звание профессора (2001). С 1998 по 2010 г. работал в СГУ имени Н. Г. Чернышевского профессором и деканом физического факультета. С 2010 года по настоящее время — заведующий кафедрой физики Саратовского государственного технического университета имени Гагарина Ю. А. Область научных интересов: оптика, лазерная физика, математическое моделирование, биофизика, применение оптических методов в биомедицинских приложениях. Имеет более 250 научных публикаций по указанным выше направлениям.

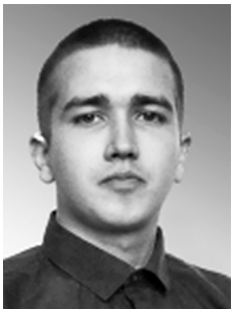
Россия, 410054 Саратов, ул. Политехническая, 77
Саратовский государственный технический университет имени Гагарина Ю. А.
E-mail: zimnykov@mail.ru
ORCID: 0000-0002-9787-7903
AuthorID (eLibrary.Ru): 29094

Алонова Марина Васильевна — окончила Саратовский государственный университет имени Н. Г. Чернышевского по специальности «Медицинская физика» (2010) и магистратуру Саратовского государственного технического университета имени Гагарина Ю. А. (СГТУ) по направлению «Приборостроение» (2012). Кандидат физико-математических наук (2019). С 2019 года доцент кафедры физики Физико-технического института СГТУ. С 2020 года заместитель директора Физико-технического института СГТУ по научной работе и ученый секретарь диссертационного совета «Математическое моделирование, численные методы и комплексы программ» при СГТУ имени Гагарина Ю. А. Область научных интересов: оптика, лазерная физика, математическое моделирование, биофизика, применение оптических методов в биомедицинских приложениях. Имеет более 50 научных публикаций по указанным выше направлениям.

Россия, 410054 Саратов, ул. Политехническая, 77
Саратовский государственный технический университет имени Гагарина Ю. А.
E-mail: alonova_marina@mail.ru
ORCID: 0000-0001-7772-3985
AuthorID (eLibrary.Ru): 667431

Скрипаль Анатолий Владимирович — окончил физический факультет Саратовского государственного университета (1982). Защитил диссертации на соискание ученой степени кандидата (1991) и доктора физико-математических наук (1998). Работает заведующим кафедрой медицинской физики в СГУ имени Н. Г. Чернышевского. Область научных интересов: лазерная физика, биотехнические системы, когерентная оптика, медицинская физика. Имеет более 400 научных публикаций по указанным выше направлениям.

Россия, 410012 Саратов, ул. Астраханская, 83
Саратовский национальный исследовательский
государственный университет имени Н. Г. Чернышевского
E-mail: skripalav@info.sgu.ru
ORCID: 0000-0002-9080-0057
AuthorID (eLibrary.Ru): 33340



Инкин Максим Глебович — окончил бакалавриат (2016) и магистратуру (2018) физического факультета Саратовского университета по направлению «Физика». Работает администратором Точки кипения в СГУ имени Н. Г. Чернышевского. Имеет более 30 научных публикаций в области статистической электроники и лазерной физики.

Россия, 410012 Саратов, ул. Астраханская, 83
Саратовский национальный исследовательский
государственный университет имени Н. Г. Чернышевского
E-mail: sunbeam18.95@mail.ru
ORCID: 0000-0002-1580-5413
AuthorID (eLibrary.Ru): 924742

Зайцев Сергей Сергеевич — окончил Саратовский государственный аграрный университет (2010). Защитил диссертацию на соискание ученой степени кандидата биологических наук (2023). Работает научным сотрудником лаборатории фундаментальных и прикладных исследований, старшим преподавателем кафедры микробиологии и биотехнологии Саратовского государственного университета генетики, биотехнологии и инженерии имени Н. И. Вавилова. Область научных интересов: генетика, биоинформатические подходы в обработке данных, молекулярная биология. Имеет более 100 научных публикаций по указанным выше направлениям.

Россия, 410012 Саратов, пр-кт им. Петра Столыпина зд. 4, стр. 3
Саратовский государственный университет генетики,
биотехнологии и инженерии имени Н. И. Вавилова
E-mail: zaytsev-sergey@inbox.ru
ORCID: 0000-0002-1373-8229
AuthorID (eLibrary.Ru): 783637

Федорова Валентина Анатольевна — окончила лечебный факультет Саратовского государственного медицинского университета (1987). Защитила диссертации на соискание ученой степени кандидата (1994) и доктора медицинских наук (2004). Имеет ученое звание профессора по специальности «Аллергология и иммунология» (2008). Работает профессором кафедры микробиологии и биотехнологии, главным научным сотрудником лаборатории фундаментальных и прикладных исследований Саратовского государственного университета генетики, биотехнологии и инженерии имени Н. И. Вавилова. Область научных интересов: возбудители инфекций, общих для человека и животных, молекулярная биология и иммунология, био- и иммуноинформатика, молекулярное моделирование. Имеет более 400 научных публикаций в профильных российских и зарубежных научных изданиях.

Россия, 410005 Саратов, ул. Соколовая, 335
Саратовский государственный университет генетики,
биотехнологии и инженерии имени Н. И. Вавилова
E-mail: feodorovav@mail.ru
ORCID: 0000-0002-3827-407X
AuthorID (eLibrary.Ru): 109625