БИОФИЗИКА И МЕДИЦИНСКАЯ ФИЗИКА

НАУЧНЫЙ ОТДЕЛ

УДК 535.41

ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ОПТИЧЕСКИХ GB-СПЕКЛОВ, ФОРМИРУЮЩИХСЯ ПРИ РАССЕЯНИИ СВЕТА НА ВИРТУАЛЬНЫХ СТРУКТУРАХ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ГЕНОВ ЭНТЕРОБАКТЕРИЙ

С. С. Ульянов, О. В. Ульянова, С. С. Зайцев, М. А. Хижнякова, Ю. В. Салтыков, Н. Н. Филонова, И. А. Субботина, А. М. Ляпина, В. А. Федорова

Ульянов Сергей Сергеевич, доктор физико-математических наук, профессор кафедры медицинской физики, Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского; ведущий научный сотрудник, Федеральный исследовательский центр вирусологии и микробиологии, филиал в Саратове, prof.sergey.ulyanov@outlook.com

Ульянова Онега Владимировна, кандидат медицинских наук, старший научный сотрудник, Федеральный исследовательский центр вирусологии и микробиологии, филиал в Саратове, ulianovaov@mail.ru

Зайцев Сергей Сергеевич, научный сотрудник, Федеральный исследовательский центр вирусологии и микробиологии, филиал в Саратове, zaytsev-sergey@inbox.ru

Хижнякова Мария Александровна, младший научный сотрудник, Федеральный исследовательский центр вирусологии и микробиологии, филиал в Саратове, khizhnyakova_mariya@mail.ru

Салтыков Юрий Владимирович, научный сотрудник, Федеральный исследовательский центр вирусологии и микробиологии, филиал в Саратове; аспирант кафедры микробиологии, биотехнологии и химии, Саратовский государственный аграрный университет имени Н. И. Вавилова, saltykov3443@mail.ru

Филонова Надежда Николаевна, младший научный сотрудник, Федеральный исследовательский центр вирусологии и микробиологии, филиал в Саратове; магистрант кафедры микробиологии, биотехнологии и химии, Саратовский государственный аграрный университет имени Н. И. Вавилова, nadejda.filonova@yandex.ru

Субботина Ирина Анатольевна, младший научный сотрудник, Федеральный исследовательский центр вирусологии и микробиологии, филиал в Саратове; магистрант кафедры микробиологии, биотехнологии и химии, Саратовский государственный аграрный университет имени Н. И. Вавилова, subbotina.irinaa@mail.ru

Ляпина Анна Михайловна, научный сотрудник, Федеральный исследовательский центр вирусологии и микробиологии, филиал в Саратове, lyapina_anna@inbox.ru

Федорова Валентина Анатольевна, доктор медицинских наук, профессор, главный научный сотрудник, Федеральный исследовательский центр вирусологии и микробиологии, филиал в Саратове; профессор кафедры микробиологии, биотехнологии и химии, Саратовский государственный аграрный университет имени Н. И. Вавилова, feodorovav@mail.ru

Представлен краткий обзор методов современной биоинформатики, основанных на использовании виртуальных оптических GB-спекл-полей (gene-based speckles) и анализе их статистических характеристик. Предложен и обсужден алгоритм преобразования нуклеотидной последовательности в двумерную GB-спекл-структуру. Проведено моделирование процессов формирования GB-спекл-структур при рассеянии когерентного света на квази-случайных поверхностях, соответствующих исходным нуклеотидным последовательностям генов, кодирующих белки семейства Omptin (*SopA, OmpP, OmpT, PgtE* и *Pla*) энтеробактерий. Исследованы статистические свойства GB-спеклов. Показана возможность выявления наличия общих мотивов указанных генов с использованием методов оптики спеклов.

Ключевые слова: последовательности нуклеотидов, GB-спеклы, референтная последовательность, дифракция когерентного света, SNP, виртуальные случайные поверхности, белки Omptin, ген.

DOI: 10.18500/1817-3020-2018-18-2-123-137

Введение

Как известно, нуклеиновые кислоты (ДНК или РНК) хранят и передают генетическую информацию в живых организмах. Структурной и функциональной единицей наследственной информации является ген. Ген представляет собой последовательность нуклеотидов в молекуле нуклеиновых кислот. Молекулы ДНК состоят из четырех типов нуклеотидов. Эти нуклеотиды содержат соответственно четыре азотистых основания, а именно аденин (A), тимин (T), гуанин (G) и цитозин (C).

Нуклеотидную последовательность можно определить, используя специальную процедуру секвенирования [1, 2], которая позволяет представить первичную структуру макромолекулы в виде линейной последовательности мономеров в текстовом формате.

Любая произвольная последовательность нуклеотидов может быть искусственно синтезирована даже в том случае, если такая последовательность ранее не существовала в природе. Этот уникальный подход был разработан в пионерских исследованиях, которые недавно были проведены в корпорации Microsoft (подразделение Microsoft Research) совместно с университетом Вашингтона (Washington University) [3, 4].

Упомянутые исследования, посвященные записи цифровой информации с использованием нуклеотидных последовательностей, чрезвычайно важны с точки зрения долговременного хранения больших баз данных. По мнению экспертов [5], в настоящее время возникла серьезная угроза возникновения кризиса хранения данных. Так, например, объем созданных в 2013 г. данных составил 3.5 ZB (1 ZB = 10^{21} байт), при этом 92% от объема всей существующей информации было сгенерировано только лишь в 2012–2013 гг. Информации о величине объемов данных, сгенерированных с 2014 г. по настоящее время, не было обнаружено авторами в открытых источниках. Ожидается [5], что к 2020 г. будет сгенерировано 40 ZB информации, обеспечить хранение которой будет в принципе невозможно, используя существующие к настоящему времени носители. Однако эта проблема может быть достаточно эффективно решена при хранении данных на ДНК-носителях. Использование ДНК для хранения информации позволяет достичь плотности хранения информации на уровне 10^9 GB/мм^3 при тысячелетних сроках хранения [4]. Так, сотрудники компании Майкрософт совместно с учеными из Вашингтонского университета сохранили в форме нуклеиновой кислоты более 200 мегабайт данных [6–10]. В частности, в запись вошли некоторые оцифрованные произведения искусства, 100 величайших литературных произведений из проекта «Гутенберг», Всеобщая декларация прав человека ООН более чем на 100 языках, база данных семян некоммерческой организации Crop Trust и клип This Too Shall Pass группы OK GO в высоком разрешении [11]. Как отмечалось в статье [10], предложенный алгоритм кодирования цифровых данных на ДНК носителе является достаточно быстродействующим, а частота появления ошибок восстановления числовых данных из нуклеотидных последовательностей весьма невелика [9].

Следует отметить, что параллельно с корпорацией Майкрософт исследования в области кодирования информации проводятся в ряде других групп. Так, ранее, в 2012 г., исследователи из Гарвардской медицинской школы зашифровали [12] полный текст книги *Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves* и сохранили его на ДНК-носителях [13]. Позднее, в 2013 г., сотрудники Европейского института биоинформатики (the European Bioinformatics Institute) сохранили несколько изображений звуковых и текстовых файлов, включая 26-секундный аудиоклип, содержащий речь Мартина Лютера Кинга «I Have a Dream». Успешная попытка сохранить на ДНК-носителе 22 МБ информации (а именно был записан немой фильм *Trip to the Moon*) и восстановить эти данные обратно в числовой формат была детально описана в статье [14].

Однако подлинный прогресс был достигнут в недавних исследованиях, опубликованных в работе [15]. Видеопоток, содержащий последовательность из пяти изображений скачущей лошади, был заархивирован в геноме живых бактерий, сохраняющих способность к размножению.

Уместно также упомянуть работы [16–19], имеющие прямое отношение к проблеме кодирования числовых данных с использованием генетических носителей информации.

В определенной степени обратная задача (а именно преобразование первичных генетических данных в числовой формат с последующей регистрацией этих данных на физическом носителе) также является чрезвычайно актуальной и перспективной. Актуальность проблемы состоит в том, что для нахождения одинаковых или идентичных фрагментов в нуклеотидной последовательности двух разных сравниваемых между собой генов требуется их последовательная обработка несколькими специальными компьютерными программами. Это занимает довольно продолжительное время и зачастую сопровождается трудностями в трактовке результатов из-за значительного количества (до 20%) ошибок на этапе секвенирования, что в настоящее время решается путем дополнительного неоднократного (иногда до 3–5 раз) ресеквенирования исходной матрицы с повторным многоступенчатым анализом. Очевидно, что используемый алгоритм вызывает неудобства и затрудняет обработку данных даже при работе с небольшими нуклеотидными последовательностями размером 250–500 нуклеотидов. Еще больше проблем возникает при работе с большими последовательности, даже если это отдельные гены размером 1000–1300 кб, которые не могут быть расшифрованы за одно прочтение, несмотря на доступность большого числа различных стратегий секвенирования протяженных фрагментов ДНК. Многочисленные ошибки прочтения и анализа преодолеваются путем неоднократного прочтения разных фрагментов гена с тем же пакетом многоступенчатых компьютерных вычислений. Понятно, что обработка даже таких больших с точки зрения биоинформатики молекул по сравнению с более короткими в 10–20 раз более трудоемко [20]. Но если нуклеотидная последовательность записывается в аналоговом формате на дифракционном оптическом элементе или на голограмме, то такой искусственный оптический элемент может быть использован при конструировании оптического процессора. Другими словами, использование спеклов может быть чрезвычайно полезным при оптической обработке нуклеотидных последовательностей в реальном времени, разработке экспресс-методов идентификации микроорганизмов, детекции таргетных генов патогенов и их типирования благодаря высокой скорости обработки данных, отсутствию потребности в последовательном использовании нескольких программ и минимизации или полному отсутствию ошибок. Высокая точность получения результата в полной мере соответствует одному из наиболее приоритетных научных кластеров, связанных с точной медициной (precise medicine), нацеленной на создание диагностических устройств нового поколения под условным названием precise medical devices. Таким образом, преобразование генетических данных в компьютерные голограммы или представление последовательности нуклеотидов в виде спекл-структуры позволит как значительно усовершенствовать, так и создать инструменты современной биоинформатики и в перспективе методы лабораторной диагностики инфекционных и неинфекционных болезней человека и животных [2].

Тем не менее, по мнению авторов данной статьи, в настоящее время междисциплинарные исследования в области когерентной оптики и молекулярной биологии фактически не проводились. Исключение составляют отдельные работы, посвященные функциональной голографии или геномной голографии (Functional Holography, Genome Holography [21]). Однако во избежание недоразумений следует особо подчеркнуть, что термин «голография» использован в статье [21] совершенно некорректно и не имеет абсолютно никакого отношения к оптической голографии.

Ранее авторами данной статьи последовательности нуклеотидов гена *omp1* бактерии *Chlamydia trachomatis* (геновары D, E, F, G, J и K) и бактерии *C. psittaci* были успешно конвертированы в двумерные спекл-поля. В работах [22–24] был введен специальный термин GB-спекл-структуры (gene-based speckles) для определения принципиально нового класса спекл-полей. GB-спекл-поля обладают уникальными статистическими свойствами, которые были частично исследованы в работе [23]. Как было показано в статье [22], использование таких методов спекл-оптики, как спекл-коррелометрия, вычитание изображений и спекл-интерферометрия, позволяет определить наличие природных мутаций в сравниваемых штаммах даже в случае минимальных различий всего в один нуклеотид

(SNP, single nucleotide polymorphism). При этом показано, что появление некоторых типов мутаций (в частности, делеций [2]) ведет к формированию полос в интерференционной картине при использовании метода спекл-интерферометрии [22]. В работе [24] проведена оптимизация алгоритма кодирования нуклеотидных последовательностей бактерии *C. trachomatis* в двумерные GB-спекл-поля, показано, что алгоритм, описанный в [22], близок к оптимальному. В статье [25] метод виртуальной спекл-интерферометрии фазового сдвига (4 bucket technique) был использован для исследования полиморфизма у двух вариантов *omp1* гена *C. trachomatis* (а именно штаммов E/Bour (E1 sub-type) и E/IU-4 2 0755u4 (E2 subtype)). Предложенный метод был успешно применен для детектирования гена *omp1 C. trachomatis* всех известных субтипов, несущих генетические мутации в виде одиночных SNP или их комбинации.

В данной статье с использованием GB-спеклов был проведен анализ нуклеотидных последовательностей генов, кодирующих продукцию сериновых протеаз, белков семейства Omptin, энтеробактерий – возбудителей таких актуальных инфекций, как сальмонелезы, иерсиниозы, шигелезы и эшерихиозы [26–28]. Сравнивались последовательности генов *pla* (*Yersinia pestis*), *pgtE* (*Salmonella enterica*), *sopA* (*Shigella flexneri*), *ompT* and *ompP* (*Escherichia coli*). Упомянутые последовательности существенно различаются между собой, но, однако, имеют общие мотивы. Под мотивом (motif) в молекулярной биологии понимается характерная относительно короткая последовательность нуклеотидов в нуклеиновых кислотах или аминокислот в полипептидах, слабо меняющаяся в процессе эволюции и имеющая определённую биологическую функцию [29]. Существование общих мотивов у последовательностей белков семейства Omptin с гомологией 40–78% была доказана на основе структурных исследований, результаты которых были ранее опубликованы в работе [30]. Недавно в результате молекулярного картирования модельного белка группы Omptin нами было продемонстрировано наличие у последних общих антигенных детерминант, имеющих потенциально диагностическое значение [31]. Показано, что экспериментальный вариант ТИФА – модификация пептидного ELISA на основе общих маркерных пептидов белков Omptin, может быть использован для ретроспективной диагностики инфекционных заболеваний, вызванных энтеропатогенными бактериальными агентами. В данной статье сходства у упомянутых последовательностей были подтверждены на основе сопоставления статистических характеристик сгенерированных GB-спеклов.

1. Преобразование последовательности нуклеотидов в спекл-структуру

На первом этапе последовательность букв из исходной одномерной нуклеотидной последовательности преобразуется в последовательность чисел в соответствии со следующим правилом [22]:

A -> 1, C -> 2, G -> 3, T -> 4. (1)

Важно подчеркнуть, что, как было показано в работе [24], взаимосвязь букв и чисел в данном случае не является принципиальной. Иными словами, при кодировании могло быть использовано любое другое правило, например:

T -> 1, G -> 2, C -> 3, A -> 4. (2)

На втором этапе генерируются все возможные комбинации (триады), содержащие лишь три числа из исходного полного набора из всех четырех чисел {1, 2, 3, 4}.

В результате формируется полный набор всех триад:

(1 1 1), (1 1 2), (1 1 3), (1 1 4), (1 2 1), (1 2 2), (1 2 3), (1 2 4), (1 3 1), ... , (4 4 4).

Число всех возможных комбинаций из четырех чисел, объединенных в триады, равно 64.

Затем на следующем (третьем) этапе некоторая дискретная величина h приписывается каждой триаде в соответствии с несложным алгоритмом, описанном в статье [22]. Упомянутый алгоритм был реализован на Matlab R2015a.

Величина h является целым числом, варьирующемся в интервале от 1 до 64. При этом каждая триада из исходной нуклеотидной последовательности ассоциируется только с одним значением h. Так, например, комбинация (1 1 1) соответствует величине h=1, (1 1 2) соответствует h = 2, (1 1 3) соответствует h=3, (1 1 4) соответствует h = 4, (1 2 1) соответствует h = 5, (1 2 2) соответствует h = 6 и так далее. Окончательно последняя комбинация (4 4 4) соответствует величине h=64.

На четвертом этапе из одномерного массива h формируется квадратная матрица $H_{n,m}$.

Физический смысл сформированной матрицы H состоит в том, что каждый ее элемент представляет собой локальную высоту некоей виртуальной шероховатой последовательности, соответствующей локальному содержанию ана-

лизируемой генетической структуры. Полученные виртуальные шероховатые поверхности будут использованы для моделирования уникальных спекл-структур, соответствующих различным специфическим нуклеотидным последовательностям.

Двумерное спекл-поле, соответствующее каждой конкретной нуклеотидной поверхности, генерируется с использованием дифракции когерентного пучка с профилем квадратного сечения на виртуальной рассеивающей поверхности с микрорельефом, описываемым матрицей $H_{n,m}$.

Как уже упоминалось, $H_{n,m}$ задает высоты шероховатости поверхности. В каждой точке виртуального диффузора (в плоскости рассеяния пучка) вводится некоторая фазовая модуляция $U_{n,m}=\exp(-2\pi i H_{n,m}/64)$. Поверхность освещается при нормальном падении пучка, фаза в освещающем пучке является постоянной величиной.

Процедура перекодирования исходной нуклеотидной последовательности в GB-спекл-структуру на примере гена *pla Y. pestis*, экспрессирующего продукцию *Pla* протеазы – типичного представителя семейства белков Omptin энтеробактерий [30], приведена ниже.

Исходная нуклеотидная последовательность *pla* (номер доступа в GenBank: AL109969.1) выглядит следующим образом:

ATGAAGAAAAGTTCTATTGTGGCAAC
CATTATAACTATTCTGTCCGGGAGTGC
TAATGCAGCATCATCTCAGTTAATAC
CAAATATATCCCCTGACAGCTTTACAGTT
GCAGCCTCCACCGGGATGCTGAGTG
GAAAGTCTCATGAAATGCTTTATGACG
CAGAAACAGGAAGAAAGATCAGCCAGT
TAGACTGGAAGATCAAAAATGTCGCTATCCT
GAAAGGTGATATATCCTGGGATCCATACT
CATTTCTGACCCTGAATGCCAGGGGGTG
GACGTCTCTGGCTTCCGGGTCAGGTAATATG
GATGACTACGACTGGATGAATGAAAAT
CAATCTGAGTGGACAGATCACTCATCT
CATCCTGCTACAAATGTTAATCATGCCAAT
GAATATGACCTCAATGTGAAAGGCTGGT
TACTCCAGGATGAGAATTATAAAGCAG
GTATAACAGCAGGATATCAGGAAACAC
GTTTCAGTTGGACAGCTACAGGTG
GTTCATATAGTTATAATAATGGAGCTTATAC
CGGAAACTTCCCGAAAGGAGTGCGGGTA
ATAGGTTATAACCAGCGCTTTTCTATGC
CATATATTGGACTTGCAGGCCAGTATCGCAT
TAATGATTTTGAGTTAAATGCATTATTTA
AATTCAGCGACTGGGTTCGGGCACAT
GATAATGATGAGCACTATATGAGAGATCT
TACTTTCCGTGAGAAGACATCCGGCTCAC
GTTATTATGGTACCGTAATTAACGCTGGATAT
TATGTCACACCTAATGCCAAAGTCTTT
GCGGAATTTACATACAGTAAATATGAT
GAGGGCAAAGGAGGTACTCAGACCATT
GATAAGAATAGTGGAGATTCTGTCTC
TATTGGCGGAGATGCTGCCGGTATTTC
CAATAAAAATTATACTGTGACGGCGGGTCT
GCAATATCGCTTCTGA

Преобразованная в числовой формат эта же последовательность принимает следующий вид:

1 4 3 1 1 3 1 1 1 1 3 4 4 2 4 1 4 4 3 4 3 3 2 1 1 2 2
1 4 4 1 4 1 1 2 4 1 4 4 2 4 3 4 2 2 3 3 3 1 3 4 3 2 4
1 1 4 3 2 1 3 2 1 4 2 1 4 2 4 2 1 3 4 4 1 1 4 1 2 2 1
1 1 4 1 4 1 4 2 2 2 2 4 3 1 2 1 3 2 4 4 4 1 2 1 3 4 4
3 2 1 3 2 2 4 2 2 1 2 2 3 3 3 1 4 3 2 4 3 1 3 4 3 3 1
1 1 3 4 2 4 2 1 4 3 1 1 1 4 3 2 4 4 4 1 4 3 1 2 3 2 1
3 1 1 1 2 1 3 3 1 1 3 1 1 1 3 1 4 2 1 3 2 2 1 3 4 4 1
3 1 2 4 3 3 1 1 3 1 4 2 1 1 1 1 1 4 3 4 2 3 2 4 1 4 2
2 4 3 1 1 1 3 3 4 3 1 4 1 4 1 4 2 2 4 3 3 3 1 4 2 2 1
4 1 2 4 2 1 4 4 4 2 4 3 1 2 2 2 4 3 1 1 4 3 2 2 1 3 3
3 3 3 4 3 3 1 2 3 4 2 4 2 4 3 3 2 4 4 2 2 3 3 3 4 2 1
3 3 4 1 1 4 1 4 3 3 1 4 3 1 2 4 1 2 3 1 2 4 3 3 1 4 3
1 1 4 3 1 1 1 1 4 2 1 1 4 2 4 3 1 3 4 3 3 1 2 1 3 1 4
2 1 2 4 2 1 4 2 4 2 1 4 2 2 4 3 2 4 1 2 1 1 1 4 3 4 4
1 1 4 2 1 4 3 2 2 1 1 4 3 1 1 4 1 4 3 1 2 2 4 2 1 1 4
3 4 3 1 1 1 3 3 2 4 3 3 4 4 1 2 4 2 2 1 3 3 1 4 3 1 3
1 1 4 4 1 4 1 1 1 3 2 1 3 3 4 1 4 1 1 2 1 3 2 1 3 3 1
4 1 4 2 1 3 3 1 1 1 2 1 2 3 4 4 4 2 1 3 4 4 3 3 1 2 1
3 2 4 1 2 1 3 3 4 3 3 4 4 2 1 4 1 4 1 3 4 4 1 4 1 1 4
1 1 4 3 3 1 3 2 4 4 1 4 1 2 2 3 3 1 1 1 2 4 4 2 2 2 3
1 1 1 3 3 1 3 4 3 2 3 3 3 4 1 1 4 1 3 3 4 4 1 4 1 1 2
2 1 3 2 3 2 4 4 4 4 2 4 1 4 3 2 2 1 4 1 4 1 4 4 3 3 1
2 4 4 3 2 1 3 3 2 2 1 3 4 1 4 2 3 2 1 4 4 1 1 4 3 1 4
4 4 4 3 1 3 4 4 1 1 1 4 3 2 1 4 4 1 4 4 4 4 1 1 1 4 4 2
1 3 2 3 1 2 4 3 3 3 4 4 2 3 3 3 2 1 2 1 4 3 1 4 1 1 4
3 1 4 3 1 3 2 1 2 4 1 4 1 4 3 1 3 1 3 1 4 2 4 4 1 2 4
4 4 2 2 3 4 3 1 3 1 1 3 1 2 1 4 2 2 3 3 2 4 2 1 2 3 4
4 1 4 4 1 4 3 3 4 1 2 2 3 4 1 1 4 4 1 1 2 3 2 4 3 3 1
4 1 4 4 1 4 3 4 2 1 2 1 2 2 4 1 1 4 3 2 2 1 1 1 3 4 2
4 4 4 3 2 3 3 1 1 4 4 4 1 2 1 4 1 2 1 3 4 1 1 1 4 1 4
3 1 4 3 1 3 3 3 2 1 1 1 3 3 1 3 3 4 1 2 4 2 1 3 1 2 2
1 4 4 3 1 4 1 1 3 1 1 4 1 3 4 3 3 1 3 1 4 4 2 4 3 4 2
4 2 4 1 4 4 3 3 2 3 3 1 3 1 4 3 2 4 3 2 2 3 3 4 1 4 4
4 2 2 1 1 4 1 1 1 1 1 4 4 1 4 1 2 4 3 4 3 1 2 3 3 2 3
3 3 4 2 4 3 2 1 1 4 1 4 2 3 2 4 4 2 4 3 1

В результате дифракции когерентного пучка с квадратным сечением формируется GB-спекл-структура (рис. 1).

Двумерное распределение фазы GB-спекл-структуры показано на рис. 2.

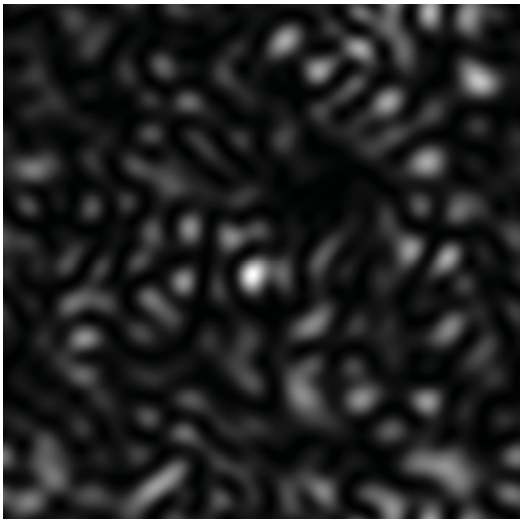


Рис. 1. Характерный вид GB-спекл-структуры (для гена *pla*)

Fig. 1. Typical view of GB-speckles (for *pla* gene)

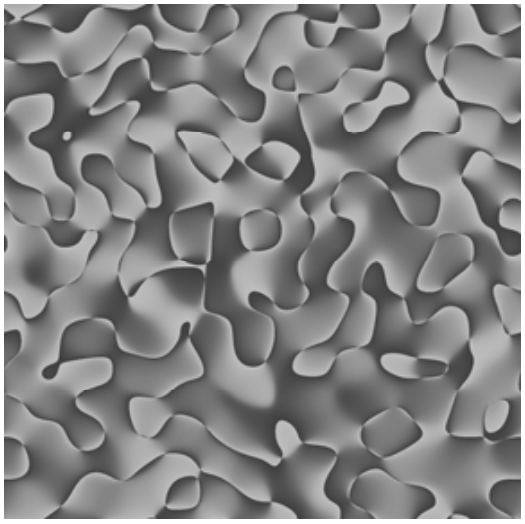


Рис. 2. Характерный вид двумерного распределения фазы GB-спекл-структуры (для гена *pla*)

Fig. 2. Typical view of the 2D phase distribution in GB-speckles (for *pla* gene)

Для сравнения на рис. 3 показана GB-спекл-структура, полученная для *C. trachomatis* геновара D, штамм B120, субтип D1. Аналогичный рисунок был уже опубликован ранее в статье [22], в данной статье он представлен в значительно большем пространственном разрешении (а именно 2048 пикселей на 2048 пикселей).

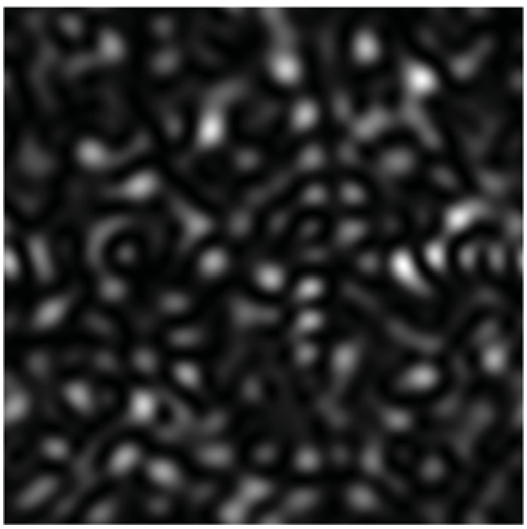


Рис. 3. GB-спекл-структура, вычисленная для *C. trachomatis* геновара D, штамм B-120, субтип D1 [22]

Fig. 3. GB-speckle-structure, computed for *C. trachomatis* genovar D, strain B-120, subtype D1 [22]

Сравнивая спекл-поля, изображенные на рис. 2 и рис. 3, даже без проведения детального статистического анализа, можно сделать заключение о принципиальных структурных различиях между GB-спеклами, полученных для различных генов.

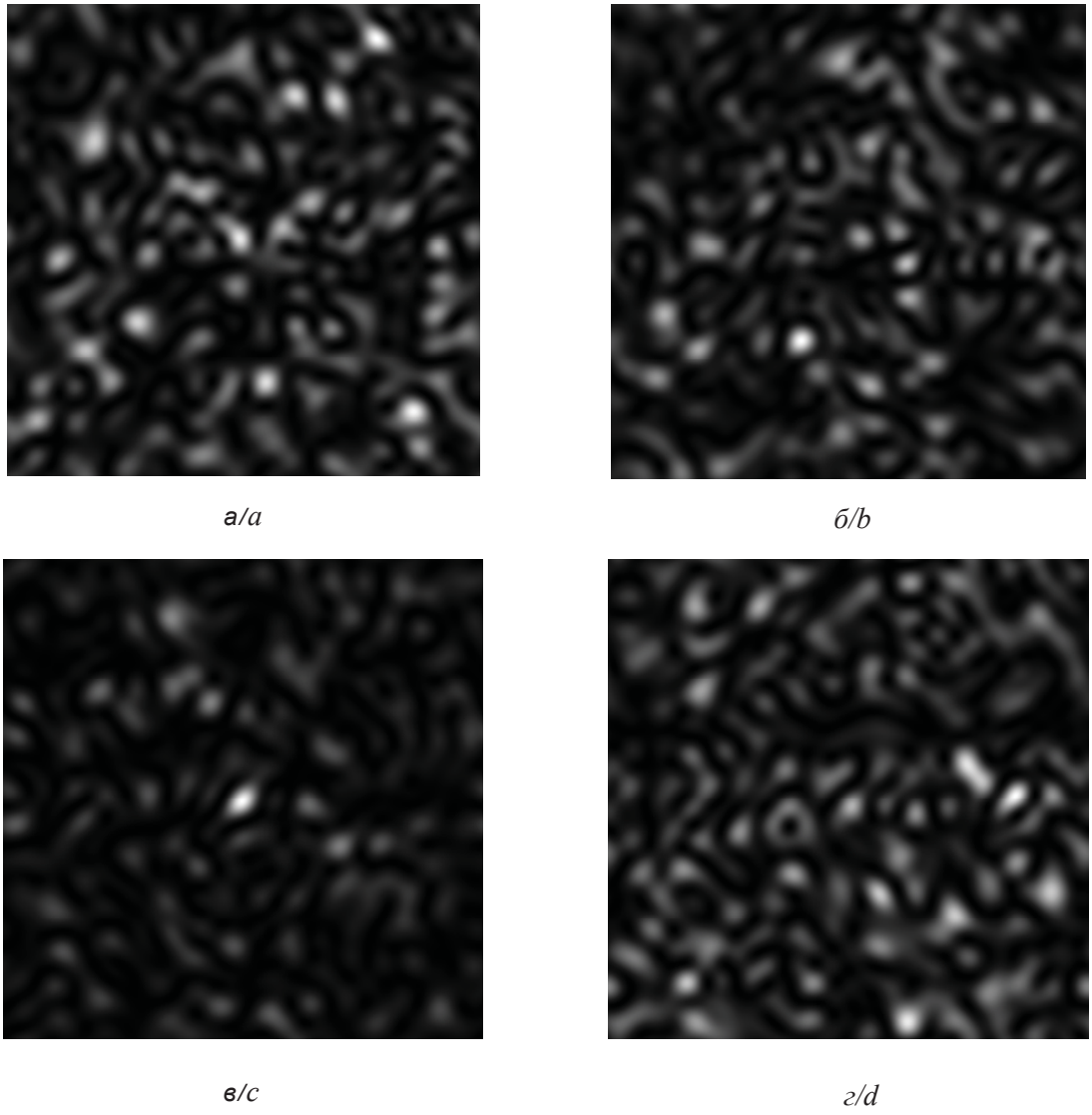
2. Моделирование GB-спекл-структур, формирующихся при рассеянии света на виртуальных квазислучайных поверхностях, полученных для нуклеотидных последовательностей генов, кодирующих биосинтез белков семейства Omptin

На рис. 4 представлены спекл-структуры, сформированные в результате дифракции лазерного пучка с прямоугольным сечением на виртуальных шероховатых поверхностях, соответствующих генам, экспрессирующим белки *SopA, OmpP, OmpT* и *PgtE* семейства Omptin.

Из сравнения рис. 1 и рис. 4, *а–г* видно, что каждому гену (*pla*, *sopA*, *ompP*, *ompT* и *pgtE*) соответствует характерная абсолютно уникальная структура GB-спекл-полей.

Двумерное распределение фазы различных GB-спеклов представлено на рис. 5.

Сопоставление рис. 2 и рис. 5, *а–г* также демонстрирует уникальность фазовой структуры GB-спеклов, соответствующих различным нуклеотидным последовательностям генов *pla*, *sopA*, *ompP*, *ompT* и *pgtE*.



а/a *б/b*

в/c *г/d*

Рис. 4. GB-спеклы, полученные для различных генов, кодирующих белки семейства Omptin: *а – SopA*, *б – OmpP*, *в – OmpT*, *г – PgtE*

Fig. 4. GB-speckles, obtained for different genes, which code the enzymes of Omptin family: *a – SopA*, *b – OmpP*, *c – OmpT*, *d – PgtE*

3. Статистические свойства GB-спекл-полей

3.1. Функции распределения плотности вероятности флуктуаций интенсивности и фазы GB-спеклов генов, кодирующих белки семейства Omptin

Для сгенерированных GB-спекл-структур, полученных для нуклеотидных последовательностей генов *pla*, *sopA*, *ompP*, *ompT* и *pgtE*, были вычислены выборочные функции распределения плотности вероятности, которые представлены на рис. 6 и рис. 7. На рис. 6 показаны распределения для пространственных флуктуаций интенсивности.

Видно, что формы функций распределения, представленные на рис. 6, близки к экспоненциальным. Гипотеза об экпоненциальности распределений была проверена с использованием критерия X^2. Как показывают результаты проверки гипотезы, пространственные флуктуации интенсивности в GB-спекл-полях подчиняются экспоненциальному распределению при уровне значимости α=0.01 для всех типов исследуемых нуклеотидных последовательностей белков семейства Omptin.

На рис. 7 представлены функции распределения плотности вероятности пространственных флуктуаций фазы в GB-спеклах.

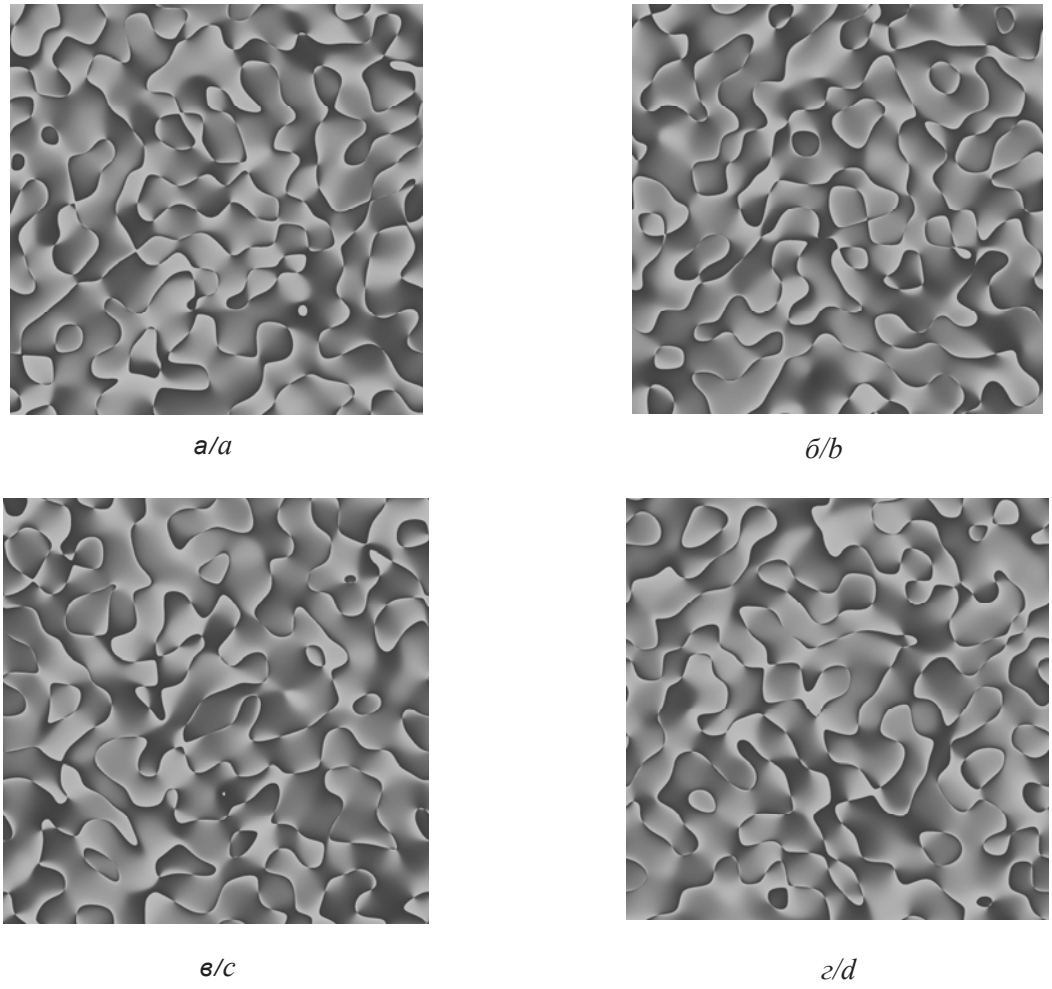


Рис. 5. Двумерное распределение фазы GB-спекл-структуры для случая дифракции лазерного излучения на виртуальных шероховатых поверхностях, соответствующих генам *sopA* (а), *ompP* (*б*), *ompT* (*в*) и *pgtE* (*г*)

Fig. 5. 2D phase distribution for GB-speckles for the case of diffraction of laser irradiation on virtual rough surfaces, corresponding to genes *sopA* (*a*), *ompP* (*b*), *ompT* (*c*), and *pgtE* (*d*)

Формы гистограмм, представленных на рис. 7, свидетельствуют о близости распределения фазы к равномерному распределению. Однако гипотеза о равномерном распределении фазы в GB-спеклах была отклонена при использовании критерия X^2 при уровне значимости $\alpha = 0.01$ для всех типов исследуемых нуклеотидных последовательностей генов, детерминирующих биосинтез белков семейства Omptin.

Однако при этом следует особо отметить, что использование критериев Шермана, критерия Морана, критерия Ченга–Спиринга, а также критерия Саркади–Косика [32] позволяет принять гипотезу о равномерности распределения фазы.

Как известно [33], экспоненциальное распределение интенсивности спеклов в сочетании с равномерным распределением фазы является признаком того, что спекл-поля подчиняются гауссовой статистике. Таким образом, можно заключить, что GB-спекл-поля являются гауссовыми для случая, если нуклеотидные последовательности относятся к генам, кодирующим белки семейства Omptin. В этой связи уместно подчеркнуть, что GB-спеклы, построенные на генах *C. trachomatis* [23], являются негауссовыми и относятся к классу пространственно-неоднородных спекл-полей, формирующихся при малом числе рассеивающих событий [34].

3.2. Корреляционные свойства GB-спекл-полей, смоделированных для генов, экспрессирующих продукцию белков семейства Omptin

Для GB-спекл-структур, представленных на рис. 1 и рис. 4, *а–г*, был проведен кросс-

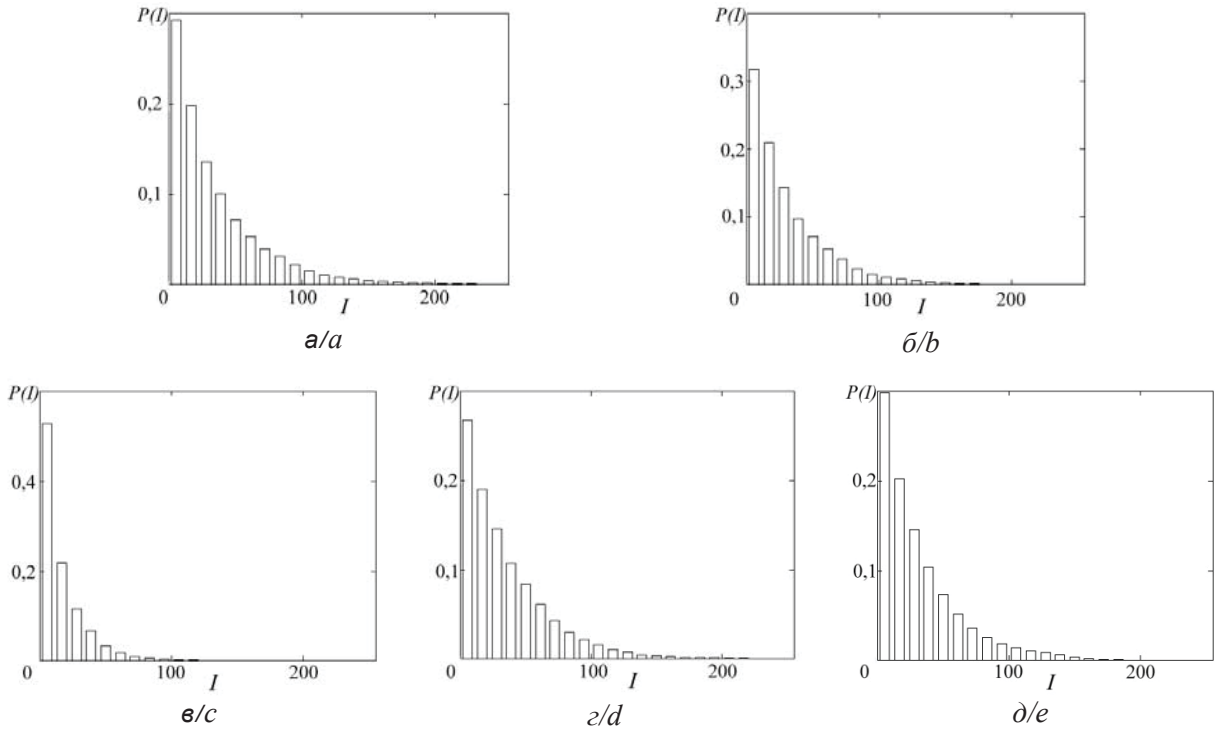


Рис. 6. Функции распределения плотности вероятности пространственных флуктуаций интенсивности GB-спеклов, полученных для последовательностей генов *pla* (а), *sopA* (*б*), *ompP* (*в*), *ompT* (*г*) и *pgtE* (*д*)

Fig. 6. Probability density functions of spatial fluctuations of the intensity in GB-speckles, obtained for gene sequences *pla* (*a*), *sopA* (*b*), *ompP* (*c*), *ompT* (*d*), and *pgtE* (*e*)

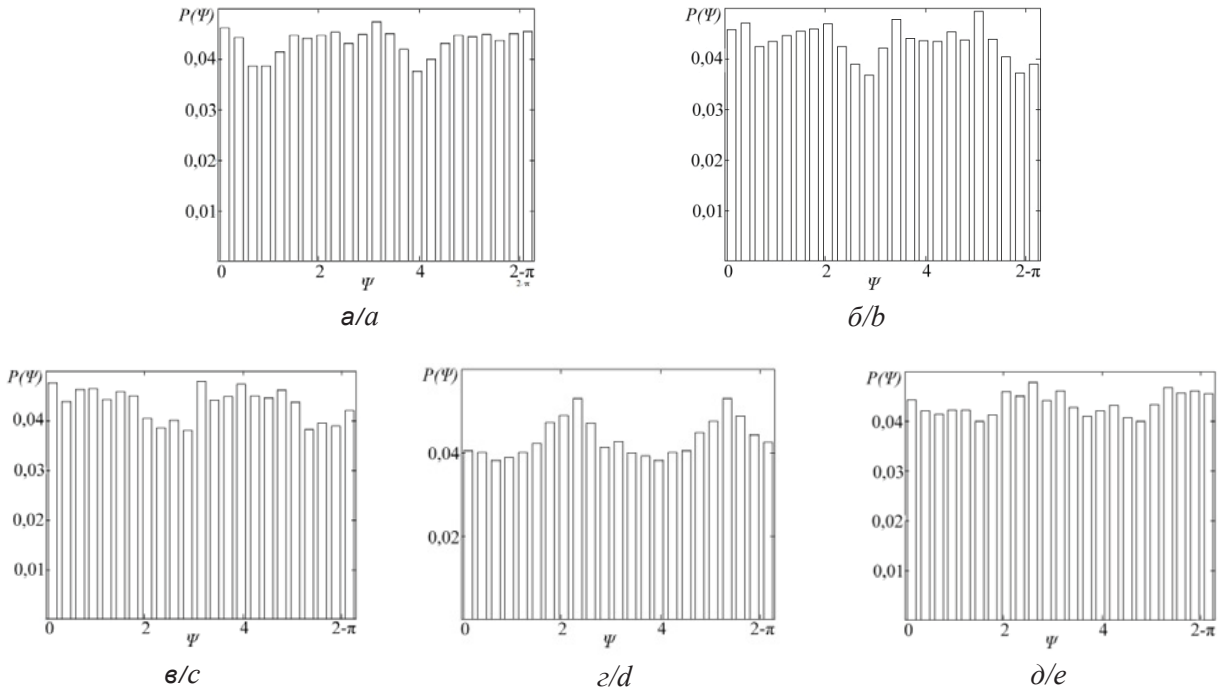


Рис. 7. Функции распределения плотности вероятности пространственных флуктуаций фазы в GB-спекл-полях, полученных для последовательностей генов *pla* (а), *sopA* (*б*), *ompP* (*в*), *ompT* (*г*) и *pgtE* (*д*)

Fig. 7. Probability density functions of spatial fluctuations of the phase in GB-speckles, obtained for gene sequences *pla* (*a*), *sopA* (*b*), *ompP* (*c*), *ompT* (*d*) and *pgtE* (*e*)

корреляционный анализ. Как показывают результаты статистических исследований, величина коэффициента кросс-корреляции (для пространственного распределения интенсивности в сопоставляемых спекл-структурах) лежит в интервале $[-3.92\times10^{-3}; 0.039]$, среднее значение коэффициента кросс-корреляции равно 0.123 при среднем квадратическом отклонении, равном 0.015. Исследования пространственного распределения фазы в спекл-структуре показывают, что величина коэффициента кросс-корреляции фазы в сравниваемых спекл-полях лежит в диапазоне [-0.011; 0.462], среднее значение коэффициента кросс-корреляции фазы равно 0.12, при среднем значении среднеквадратического отклонения коэффициента кросс-корреляции фазы, равного 0.153. То есть корреляция между GB-спекл-структурами и соответствующими генами, имеющими общие мотивы, крайне мала.

Еще одной важной характеристикой, определяющей меру сходства двух сравниваемых двумерных полей, является расстояние Хэмминга (см., например, [35]).

В данной статье используется величина, являющаяся аналогом расстояния Хемминга, вычисляемая по следующей формуле:

$$HD = \frac{M(I1_{i,j} = = I2_{(i,j)})}{M_{total}}, \qquad (3)$$

где $I1_{i,j}$ и $I2_{(i,j)}$ – интенсивности в двух сравниваемых спекл-структура; i и j – номера строки и ряда для каждого пикселя спекл-структуры; M – количество пикселей, в которых значения интенсивности в первом спекл-поле в точности совпадает со значением интенсивности во втором спекл-поле (т.е. $I1_{i,j} = = I2_{(i,j)}$); M_{total} – полное количество пикселей в каждой спекл-структуре. В рассматриваемом случае M_{total} всегда равно 4.192×10^{6}. Очевидно, что при абсолютной схожести сравниваемых двумерных полей HD принимает максимально возможное значение, равное 1. Если интенсивность света в сопоставляемых полях не имеет одинаковых значений ни в одной точке, то $HD = 0$.

Как показывают результаты компьютерных вычислений, величина HD для пространственного распределения интенсивности в спекл-структуре лежит в интервале [0.014; 0.021], среднее значение $<HD>$ равно 0.018 при среднем квадратическом отклонении HD, равном 2.6×10^{-3}. Исследования пространственного распределения фазы в спекл-структуре показывают, что величина HD для сопоставляемых спекл-полей распределена в диапазоне $[3.721 \times 10^{-3}; 0.013]$, среднее значение HD равно 5.993×10^{-3} при среднем квадратическом отклонении HD, равном 2.608×10^{-3}.

Однако, если в качестве референтной последовательности взять GB-спекл-структуру, соответствующую генетической последовательности гена *omp1 C. trachomatis*, то значения HD будут лежать в интервале [0.014; 0.019] при среднем значении $<HD>$=0.016 и среднеквадратическом отклонении, равном 1.72×10^{-3}. Это, в свою очередь, свидетельствует о том, что столь малые величины HD появляются также и при сопоставлении GB-спекл-структур для генов, не имеющих общих мотивов.

Это означает, что, несмотря на доказанное существование общих мотивов у сравниваемых нуклеотидных последовательностей генов, кодирующих белки Omptin, построенные на их основе спекл-поля являются совершенно несхожими по своей структуре. Степень схожести GB-спеклов для генов, имеющих общие мотивы, и для генов, не имеющих общих мотивов, одинакова мала.

4. Выявление наличия общих мотивов у генов *sopA*, *ompP*, *ompT* и *pgtE* методами оптики спеклов

Проведенный в парагр. 3 анализ показывает, что использование классических методов статистики первого и второго порядка является малоэффективным с точки зрения выявления характерных особенностей GB-спекл-структур, соответствующих различным нуклеотидным последовательностям, имеющим общие мотивы. Однако существование общих мотивов в GB-спеклах может быть выявлено чрезвычайно легко при сопоставлении функций распределения плотности вероятности. При этом информативной характеристикой является следующая величина:

$$DFR(I) = P1(I) - P2(I), \qquad (4)$$

где $P1(I)$ и $P2(I)$ – по-прежнему функции распределения плотности вероятности флуктуаций интенсивности в сравниваемых спекл-полях.

Следует, однако, отметить важную особенность вычисления величины DFR. Если при построении гистограмм, приведенных на рис. 6, 7, число интервалов в ранжированном интервальном ряду вычислялось в соответствии с формулой Стержесса [36] и полагалось равным 23, то число интервалов N в формуле (4) принимало значительно большее значение. В проведенных исследованиях N было равно

2048 для спекл-структур, содержащих 2048 пикселей × 2048 пикселей. Разностные значения для сопоставляемых функций распределения плотности вероятностей (функция *DFR*(*I*)) при большом количестве интервалов показано на рис. 8. Видно, что поведение функции *DFR*(*I*) носит монотонный характер и имеется только один ноль функции. Появление нуля функции *DFR(I)* свидетельствует о существовании общих мотивов в сопоставляемых нуклеотидных последовательностях, на которых были построены GB-спекл-поля.

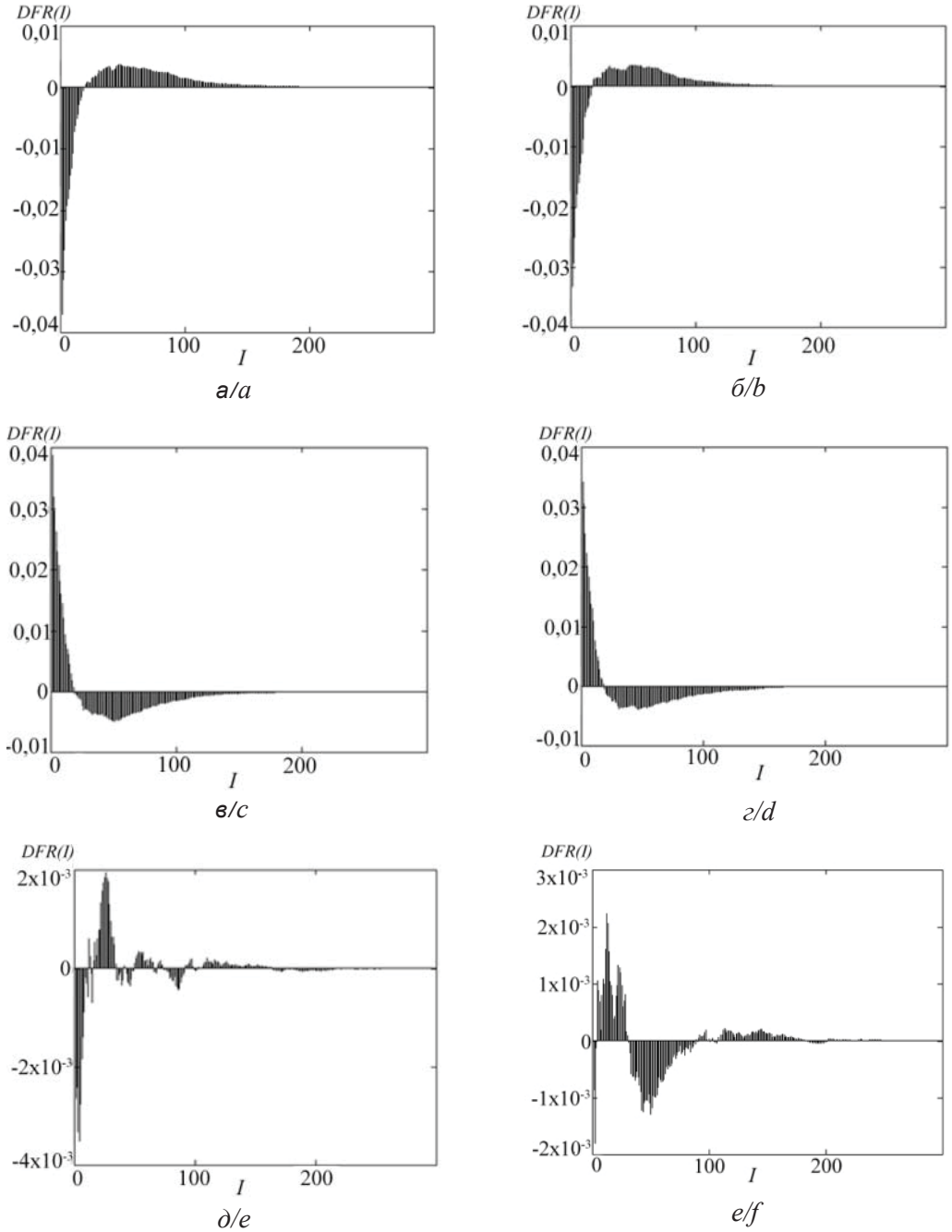


Рис. 8. График функции *DFR(I)* для различных GB-спеклов отдельных пар нуклеотидных последовательностей генов: *а – sopA–ompT, б – ompP–ompT, в – ompT–pgtE, г – ompT–pla, д – sopA–omp1 C. trachomatis, е – pgtE–omp1 C. trachomatis*

Fig. 8. Plot of function *DFR(I)* for different GB-speckles of separate pairs of nucleotide gene sequences: *a – sopA–ompT, b – ompP–ompT, c – ompT–pgtE, d – ompT–pla, e – sopA–omp1 C. trachomatis, f – pgtE–omp1 C. trachomatis*

Принципиально важным является то обстоятельство, что нули у функции $DFR(I)$ не появляются для GB-спеклов, полученных для последовательностей нуклеотидов, не имеющих общих мотивов. Так, на рис. 8, *д* и рис. 8, *е* показана функция $DFR(I)$ для спеклов, полученных для случая сопоставления генов *omp1 C. trachomatis* и *sopA* (рис. 8, *д*), а также *omp1 C. trachomatis* и *pgtE* (рис. 8, *е*) соответственно. Детальный анализ показывает, что для данного случая (см. рис. 8, *д*, рис. 8, *е*) функция $DFR(I)$ может принимать малые значения, но никогда не обращается в ноль и при этом носит немонотонный (и даже случайный) характер.

Заключение

В данной работе проведен статистический анализ GB-спекл-структур, соответствующих различным последовательностям генов, кодирующих биосинтез белков семейства Omptin. Показано, что данные спекл-поля являются гауссовыми. Установлено, что классические методы статистического анализа GB-спеклов являются неинформативными и малоэффективными с точки зрения выявления сходных фрагментов в исходных нуклеотидных последовательностях. Однако прямое сравнение функций распределения плотности вероятностей пространственных флуктуаций интенсивности спеклов позволяет достоверно выявлять общие мотивы сравниваемых генов. Критерием наличия общих мотивов является появление нуля функции $DFR(I)$ и ее монотонный характер поведения.

Благодарности

Работа выполнена при финансовой поддержке ПФНИ РАН по направлению 160, тема № 0615-2018-0001 на 2018 год.

Список литературы

1. *Sintchenko V., Roper M. P.* Pathogen genome bioinformatics // Methods in Molecular Biology. 2014. Vol. 1168. P. 173–193.
2. *Lesk A. M.* Introduction to bioinformatics. Oxford : Oxford University Press, 2002. 314 p.
3. *Guo Q., Strauss K., Ceze L., Malvar H.* High-density image storage using approximate memory cells // Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems. ASPLOS ‘16 (Apr 2–6, 2016). Atlanta (US) : IEEE, 2016. P. 413–426. DOI: http://dl.acm.org/citation.cfm?doid=2872362.2872413
4. *Bornholt J., Lopez R., Carmean D. M., Ceze L., Seelig G., Strauss K.* A DNA-based archival storage system // Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems. ASPLOS ‘16 (Apr 2–6, 2016). Atlanta (US) : IEEE, 2016. P. 637–649. DOI: http://dx.doi.org/10.1145/2872362.2872397
5. Rocky Pimentel. Why Data Storage Is Hot Again. URL: https://www.recode.net/2014/1/10/11622168/stuffed-why-data-storage-is-hot-again-really (дата обращения: 10.01.2014).
6. *Bornholt J., Lopez R., Carmean M. D.* Toward a DNA-based archival storage system // IEEE MICRO. 2017. Vol. 37, iss. 3. P. 98–104.
7. *Bornholt J., Lopez R., Ceze L.* A DNA-Based Archival Storage System // Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems. ASPLOS ‘16 (Apr 2–6, 2016). Atlanta (US) : IEEE, 2016. P. 637–649. DOI: http://dx.doi.org/10.1145/2872362.2872397
8. *Guo Q., Strauss K., Ceze L.* High-Density Image Storage Using Approximate Memory Cells. ASPLOS ‘16. 2016. P. 1–14. DOI: http://dx.doi.org/10.1145/2872362.2872413
9. *Organick L., Dumas S., Ang S. D., Chen Y.-J., Lopez R.* Scaling up DNA data storage and random access retrieval // BioRxiv. 2017. Posted March 7. P. 1–14. DOI: http://dx.doi.org/10.1101/114553
10. *Rashtchian C., Makarychev K., Rácz M.* Clustering Billions of Reads for DNA Data Storage // 31st Conference on Neural Information Processing Systems (Dec 4–9, 2017). Long Beach (US) : NIPS, 2017. P. 1–12.
11. Клип This Too Shall Pass группы OK Go. URL: https://www.youtube.com/watch?v=qybUFnY7Y8w (дата обращения: 1.03.2018).
12. *Rojahn S. Y.* An Entire Book Written in DNA. URL: https://www.technologyreview.com/s/428922/an-entire-book-written-in-dna (дата обращения: 16.08.2012).
13. *Church G. M., Gao Y., Kosuri S.* Next-Generation Digital Information Storage in DNA // Science. 2012. Vol. 337. P. 1628.
14. *Blawat M., Gaedke K., Hütter I.* Forward Error Correction for DNA Data Storage // Procedia Computer Science. 2016. Vol. 80. P. 1011–1022.
15. *Shipman S. L., Nivala J., Macklis J. D.* CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria // Nature. 2017. Vol. 547. P. 345–349. DOI: 10.1038/nature23017
16. *Goldman N., Bertone P., Chen S., Dessimoz Ch., LeProust E. M., Sipos B., Birney E.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA // Nature. 2013. Vol. 494. P. 77–80. DOI: 10.1038/nature11875
17. *Gibson D. G., Glass J. I., Lartigue C., Noskov V. N., Chuang R. Y., Algire M. A., Benders G. A., Montague M. G., Ma L., Moodie M. M., Merryman C., Vashee S.,*

Krishnakumar R., Assad-Garcia N., Andrews-Pfannkoch C., Denisova E. A., Young L., Qi Z. Q., Segall-Shapiro T. H., Calvey C. H., Parmar P. P., Hutchison C. A. 3rd., Smith H. O., Venter J. C. Creation of a bacterial cell controlled by a chemically synthesized genome // Science. 2010. Vol. 329. P. 52–56.

18. *Clelland C. T., Risca V., Bancroft C.* Hiding messages in DNA microdots // Nature. 1999. Vol. 399. P. 533–534.
19. *Adleman L. M.* Molecular computation of solutions to combinatorial problems // Science. 1994. Vol. 266. P. 1021–1024.
20. *Mandeles S.* Nucleic acid sequence analysis. N.Y. ; L. : Columbia University Press, 1972. 282 p.
21. *Madi A., Friedman Y., Roth D., Regev T., Bransburg-Zabary S., Jacob E. B.* Genome holography : Deciphering function-form motifs from gene expression data // PLoS One. 2008. Vol. 3, iss. 7. P. 114. DOI: http://dx.doi.org/10.1371/journal.pone.0002708
22. *Ulyanov S. S., Zaytsev S. S., Ulianova O. V., Saltykov Y. V., Feodorova V. A.* Using of methods of speckle optics for *Chlamydia trachomatis* typing // Proc. SPIE. 2017. Vol. 10336. Paper 03360D. DOI: 10.1117/12.2270760
23. *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 // Laser Physics Letters. 2018. Vol. 15, № 4. P. 1–6. DOI: https://doi.org/10.1088/1612-202X/aaa11c
24. *Feodorova V. A., Ulyanov S. S., Zaytsev S. S., Saltykov Y. V., Ulianova O. V.* Optimization of algorithm of coding of genetic information of Chlamydia // Proc. SPIE. 2018.Vol. 10716. Paper 107160Q. DOI: 10.1117/12.2314640
25. *Feodorova V. A., Saltykov Y. V., Zaytsev S. S., Ulyanov S. S., Ulianova O. V.* Application of virtual phase-shifting speckle-interferometry for detection of polymorphism in the *Chlamydia trachomatis omp1* gene // Proc. SPIE. 2018. Vol. 10716. Paper 107160M. DOI: 10.1117/12.2314700
26. World Health Organization. Media centre. Diarrhoeal disease. URL: http://www.who.int/mediacentre/factsheets/fs330/en/ (дата обращения: 01.05.2017).
27. *Забокрицкий Н. А.* Инфекционная заболеваемость в Российской Федерации и тенденции ее развития в ближайшее десятилетие // Вестн. «Здоровье и образование в XXI веке». 2015. Т. 17, № 5. С. 16–26.
28. Роспотребнадзор обнародовал статистику инфекционных болезней за первое полугодие. URL: http://www.yaprivit.ru/news/2365/ (дата обращения: 18.07.2016).
29. *D'haeseleer P.* What are DNA sequence motifs? // Nature Biotechnology. 2006. Vol. 24, iss. 4. P. 423–425. DOI: 10.1038/nbt0406-423
30. *Kukkonen M., Korhonen K.* The omptin family of enterobacterial surface proteases/adhesins : from housekeeping in Escherichia coli to systemic spread of Yersinia pestis // Intern. J. Med. Microbiol. 2004. July. Vol. 294, № 1. P. 7–14. DOI: https://doi.org/10.1016/j.ijmm.2004.01.003
31. *Федорова В. А., Хижнякова М. А., Зайцев С. С., Ляпина А. М., Саяпина Л. В., Ляпина Е. П., Ульянова О. В., Мотин В. Л.* Изучение диагностической значимости иммунореактивных эпитопов протеаз семейства Omptin с использованием пептидной библиотеки // Биопрепараты. 2017. Т. 17, № 3. С. 180–186.
32. *Кобзарь А. И.* Прикладная математическая статистика. Для инженеров и научных работников. М. : ФИЗМАТЛИТ, 2006. 816 с.
33. *Гудман Дж.* Статистическая оптика. М. : Мир, 1988. 528 с.
34. *Jakeman E.* Speckle statistics with a small number of scatterers // Optical Engineering. 1984. Vol. 23, № 4. P. 453–661. DOI: 10.1117/12.7973317
35. *Daugman J.* How Iris Recognition Works // IEEE transactions on circuits and systems for video technology. 2004. Vol. 14, № 1. January. P. 21–30. DOI: 10.1109/TCSVT.2003.818350
36. *Wendy L. Martinez, Angel R. Martinez.* Computational statistics handbook with Matlab. Boca Raton ; London ; N.Y. ; Washington. D.C. : Chapman & Hall / CRC, 2002. 585 p.

Образец для цитирования:

Ульянов С. С., Ульянова О. В., Зайцев С. С., Хижнякова М. А., Салтыков Ю. В., Филонова Н. Н., Субботина И. А., Ляпина А. М., Федорова В. А. Исследование статистических характеристик оптических GB-спеклов, формирующихся при рассеянии света на виртуальных структурах нуклеотидных последовательностей генов энтеробактерий// Изв. Сарат. ун-та. Нов. сер. Сер. Физика. 2018. Т. 18, вып. 2. С. 123–137. DOI: 10.18500/1817-3020-2018-18-2-123-137.

Study of Statistical Characteristics of GB-speckles, Forming at Scattering of Light on Virtual Structures of Nucleotide Gene Sequences of Enterobacteria

S. S. Ulyanov, O. V. Ulianova, S. S. Zaitsev, M. A. Khizhnyakova Yu. V. Saltykov, N. N. Filonova, I. A. Subbotina, A. M. Lyapina, V. A. Feodorova

Sergey S. Ulyanov, ORCID 0000-0003-3030-6927, Saratov State University, 83, Astrakhanskaya Str., Saratov, 410012, Russia

Onega V. Ulianova, ORCID 0000-0003-4420-2408, Federal Research Center for Virology and Microbiology, Branch in Saratov, 24, Strelkovaja Divizija Str., Saratov, 410028, Russia

Sergey S. Zaitsev, ORCID 0000-0002-1373-8229, Federal Research Center for Virology and Microbiology, Branch in Saratov, 24, Strelkovaja Divizija Str., Saratov, 410028, Russia

Yury V. Saltykov, ORCID 0000-0001-8163-1979, Federal Research Center for Virology and Microbiology, Branch in Saratov, 24, Strelkovaja Divizija Str., Saratov, 410028, Russia

Maria A. Khizhnyakova, ORCID 0000-0001-7053-8322, Federal Research Center for Virology and Microbiology, Branch in Saratov, 24, Strelkovaja Divizija Str., Saratov, 410028, Russia

Nadezhda N. Filonova, ORCID 0000-0003-1313-6241, Federal Research Center for Virology and Microbiology, Branch in Saratov, 24, Strelkovaja Divizija Str., Saratov, 410028, Russia

Irina A. Subbotina, ORCID 0000-0003-4386-5058, Federal Research Center for Virology and Microbiology, Branch in Saratov, 24, Strelkovaja Divizija Str., Saratov, 410028, Russia

Anna M. Lyapina, ORCID 0000-0003-3527-2076, researcher, Federal Research Center for Virology and Microbiology, Branch in Saratov, 24, Strelkovaja Divizija Str., Saratov, 410028, Russia

Valentina A. Feodorova, ORCID 0000-0002-3827-407X, Federal Research Center for Virology and Microbiology, Branch in Saratov, 24, Strelkovaja Divizija Str., Saratov, 410028, Russia

Background and Objectives: A brief review of methods of modern bioinformatics, based on the usage of virtual optical GB-speckles (gene-based speckles), has been presented in this paper. An algorithm of transformation of a nucleotide sequence into a 2D GB-speckle-structure has been proposed and discussed. **Materials and Methods:** Computer simulation of the process of formation of GB-speckles at the scattering of coherent light on quasi-random virtual surfaces, corresponding to initial nuclear sequence of the genes, encoded by the Omptin family proteins, such as *SopA, OmpP, OmpT, PgtE* and *Pla* in Enterobacteriaceae spp. has been carried out. **Results:** Statistical properties of GB-speckles, coding of different sequences of the genes have been investigated. **Conclusion:** It has been shown that GB-speckles of this type obey Gaussian statistics. It has also been found that classical methods of statistical analysis of GB speckles are not informative and low-effective from a viewpoint of detection of common fragments in initial nucleotide sequences. However, a direct comparison of the probability density functions of spatial fluctuations of the speckle intensity allows to find common motifs of the comparing genes. A criterion for the reliable detection of the presence of common motifs in these genes, based on the methods of speckle-optics has been suggested. These motifs could be innovated promising molecular targets for the development of a new generation of effective synthetic Omptin-based peptide precise medical devices for smart laboratory diagnostics of a group of Gram-negative Enterobacterial pathogens.

Key words: nucleotide sequence, GB speckles, reference sequence, diffraction of coherent light, SNP, virtual random surfaces, Omptin proteins, gene.

Acknowledgements: *This work was supported by PFNI RAS on 2018 year, direction 160, project no. 0615-2018-0001.*

References

1. Sintchenko V., Roper M. P. Pathogen genome bioinformatics. *Methodsin Molecular Biology*, 2014, vol.1168, pp. 173–193.
2. Lesk A. M. *Introduction to bioinformatics*. Oxford, Oxford University Press, 2002. 314 p.
3. Guo Q., Strauss K., Ceze L., Malvar H. High-density image storage using approximate memory cells. *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems. ASPLOS'16*, Apr 2–6, 2016. Atlanta (US), IEEE, 2016, pp. 413–426. DOI: http://dl.acm.org/citation.cfm?doid=2872362.2872413
4. Bornholt J., Lopez R., Carmean D. M., Ceze L., Seelig G., Strauss K. A DNA-based archival storage system. *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems. ASPLOS'16*, Apr 2–6, 2016. Atlanta (US), IEEE, 2016, pp. 637–649. DOI: http://dx.doi.org/10.1145/2872362.2872397
5. Rocky Pimentel. *Why Data Storage Is Hot Again*. Available at: https://www.recode.net/2014/1/10/11622168/stuffed-why-data-storage-is-hot-again-really (accessed 10 January 2014).
6. Bornholt J., Lopez R., Carmean M. D. Toward a DNA-based archival storage system. *IEEE MICRO*, 2017, vol. 37, iss. 3, pp. 98–104.
7. Bornholt J., Lopez R., Ceze L. A DNA-Based Archival Storage System. *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems. ASPLOS'16*, Apr 2–6, 2016. Atlanta (US), IEEE, 2016, pp. 637–649. DOI: http://dx.doi.org/10.1145/2872362.2872397
8. Guo Q., Strauss K., Ceze L. High-Density Image Storage Using Approximate Memory Cells. *ASPLOS '16*, 2016, pp. 1–14. DOI: http://dx.doi.org/10.1145/2872362.2872413
9. Organick L., Dumas S., Ang S. D., Chen Y.-J., Lopez R. Scaling up DNA data storage and random access retrieval. *BioRxiv.*, 2017, Posted March 7, pp. 1–14. DOI: http://dx.doi.org/10.1101/114553
10. Rashtchian C., Makarychev K., Rácz M. Clustering Billions of Reads for DNA Data Storage. *31st Conference on Neural Information Processing Systems*, Dec 4–9, 2017. Long Beach (US), NIPS. 2017, pp. 1–12.
11. *Video This Too Shall Pass by group OK Go*. Available at: https://www.youtube.com/watch?v=qybUFnY7Y8w (accessed 1 March 2018).
12. Rojahn S. Y. *An Entire Book Written in DNA*. Available at: https://www.technologyreview.com/s/428922/an-entire-book-written-in-dna (accessed 16 August 2012).
13. Church G. M., Gao Y., Kosuri S. Next-Generation Digital Information Storage in DNA. *Science*, 2012, vol. 337, pp. 1628.
14. Blawat M., Gaedke K., Hütter I. Forward Error Correction for DNA Data Storage. *Procedia Computer Science*, 2016, vol. 80, pp. 1011–1022.
15. Shipman S. L., Nivala J., Macklis J. D. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature*, 2017, vol. 547, pp. 345–349. DOI: 10.1038/nature23017

16. Goldman N., Bertone P., Chen S., Dessimoz Ch., LeProust E. M., Sipos B., Birney E. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 2013, vol. 494, pp. 77–80. DOI: 10.1038/nature11875
17. Gibson D. G., Glass J. I., Lartigue C., Noskov V. N., Chuang R. Y., Algire M. A., Benders G. A., Montague M. G., Ma L., Moodie M. M., Merryman C., Vashee S., Krishnakumar R., Assad-Garcia N., Andrews-Pfannkoch C., Denisova E. A., Young L., Qi Z. Q., Segall-Shapiro T. H., Calvey C. H., Parmar P. P., Hutchison C. A. 3rd., Smith H. O., Venter J. C. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 2010, vol. 329, pp. 52–56.
18. Clelland C. T., Risca V., Bancroft C. Hiding messages in DNA microdots. *Nature*, 1999, vol. 399, pp. 533–534.
19. Adleman L. M. Molecular computation of solutions to combinatorial problems. *Science*, 1994, vol. 266, pp. 1021–1024.
20. Mandeles S. *Nucleic acid sequence analysis*. New York, London, Columbia University Press, 1972. 282 p.
21. Madi A., Friedman Y., Roth D., Regev T., Bransburg-Zabary S., Jacob E. B. Genome holography: Deciphering function-form motifs from gene expression data. *PLoS One*, 2008, vol. 3, iss. 7, pp. 114. DOI: http://dx.doi.org/10.1371/journal.pone.0002708
22. Ulyanov S. S., Zaytsev S. S., Ulianova O. V., Saltykov Y. V., Feodorova V. A. Using of methods of speckle optics for *Chlamydia trachomatis* typing. *Proc. SPIE*, 2017, vol. 10336, paper 03360D. DOI: 10.1117/12.2270760
23. 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. *Laser Physics Letters*, 2018, vol. 15, no. 4, pp. 1–6. DOI: https://doi.org/10.1088/1612-202X/aaa11c
24. Feodorova V. A., Ulyanov S. S., Zaytsev S. S., Saltykov Y. V., Ulianova O. V. Optimization of algorithm of coding of genetic information of Chlamydia. *Proc. SPIE*, 2018, vol. 10716, paper 107160Q. DOI: 10.1117/12.2314640
25. Feodorova V. A., Saltykov Y. V., Zaytsev S. S., Ulyanov S. S., Ulianova O. V. Application of virtual phase-shifting speckle-interferometry for detection of polymorphism in the *Chlamydia trachomatis omp1* gene. *Proc. SPIE*, 2018, vol. 10716, paper 107160M. DOI: 10.1117/12.2314700
26. *World Health Organization. Media centre. Diarrhoeal disease*. Available at: http://www.who.int/mediacentre/factsheets/fs330/en/ Updated (accessed 1 May 2018).
27. Zabokritskiy N. A. The infectious morbidity in the Russian Federation and tendencies of it is development in the next decade. *Bulletin "Health & education millennium"'*, 2015, vol. 17, no. 5, pp. 16–26 (in Russian).
28. *Rospotrebnadzor obnarodoval statistiku infektsionnykh boleznei za pervoe polugodie* (Rospotrebnadzor published statistical data regarding inflection diseases for first half of year). Available at: http://www.yaprivit.ru/news/2365/ (accessed 18 July 2016) (in Russian).
29. D'haeseleer P. What are DNA sequence motifs? *Nature Biotechnology*, 2006, vol. 24, iss. 4, pp. 423–425. DOI: 10.1038/nbt0406-423
30. Kukkonen M., Korhonen K. The omptin family of enterobacterial surface proteases/adhesins: from housekeeping in Escherichia coli to systemic spread of Yersinia pestis. *Intern. J. Med. Microbiol.*, 2004, July, vol. 294, no. 1, pp. 7–14. DOI: https://doi.org/10.1016/j.ijmm.2004.01.003
31. Feodorova V. A., Khizhnyakova M. A., Zaitsev S. S., Lyapina A. M., Sayapina L.V., Lyapina E.P., Ulyanova O. V., Motin V. L. Evaluation of diagnostic potential of immunoreactive epitopes of the Omptin protease family by using a peptide library. *Biopreparations*, 2017, vol. 17, no. 3, pp. 180–186 (in Russian).
32. Kobzar' A. I. *Prikladnaia matematicheskaia statistika. Dlia inzhenerov i nauchnykh rabotnikov* [Applied Statistics for Engineers and Scientists]. Moscow, FIZMATLIT Publ., 2006. 816 p. (in Russian).
33. Goodman J. *Statisticheskaia optika* [Statistical Optics]. Moscow, Mir Publ., 1988. 528 p. (in Russian).
34. Jakeman E. Speckle statistics with a small number of scatterers. *Optical Engineering*, 1984, vol. 23, no. 4, pp. 453–661. DOI:10.1117/12.7973317
35. Daugman J. How Iris Recognition Works. *IEEE transactions on circuits and systems for video technology*, 2004, vol. 14, no. 1, January, pp. 21–30. DOI: 10.1109/TCSVT.2003.818350
36. Wendy L. Martinez., Angel R. Martinez. *Computational statistics handbook with Matlab.* Boca Raton, London, New York, Washington, D.C., Chapman & Hall / CRC, 2002. 585 p.

Cite this article as:

Ulyanov S. S., Ulianova O. V., Zaitsev S. S., Khizhnyakova M. A., Saltykov Yu. V., Filonova N. N., Subbotina I. A., Lyapina A. M., Feodorova V. A. Study of Statistical Characteristics of GB-speckles, Forming at Scattering of Light on Virtual Structures of Nucleotide Gene Sequences of Enterobacteria. *Izv. Saratov Univ. (N. S.), Ser. Physics*, 2018, vol. 18, iss. 2, pp. 123–137 (in Russian). DOI: 10.18500/1817-3020-2018-18-2-123-137.