**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ**
**И МОДЕЛИРОВАНИЕ 2020 Т. 12 № 2 С. 417–434**
DOI: 10.20537/2076-7633-2020-12-2-417-434

Ки&М

**АНАЛИЗ И МОДЕЛИРОВАНИЕ СЛОЖНЫХ ЖИВЫХ СИСТЕМ**

УДК: 519.688, 51-76

# Биоматематическая система методов описания нуклеиновых кислот

**И. В. Степанян**

Институт машиноведения им. А. А. Благонравова Российской академии наук,
Россия, 101990, г. Москва, Малый Харитоньевский переулок, д. 4

E-mail: neurocomp.pro@gmail.com

*Получено 11.10.2019, после доработки — 17.12.2019.*
*Принято к публикации 26.12.2019.*

Статья посвящена применению методов математического анализа, поиска паттернов и изучения состава нуклеотидов в последовательностях ДНК на геномном уровне. Изложены новые методы математической биологии, которые позволили обнаружить и отобразить скрытую упорядоченность генетических нуклеотидных последовательностей, находящихся в клетках живых организмов. Исследования основаны на работах по алгебраической биологии доктора физико-математических наук С. В. Петухова, которым впервые были введены и обоснованы новые алгебры и гиперкомплексные числовые системы, описывающие генетические явления. В данной работе описана новая фаза развития матричных методов в генетике для исследования свойств нуклеотидных последовательностей (и их физико-химических параметров), построенная на принципах конечной геометрии. Целью исследования является демонстрация возможностей новых алгоритмов и обсуждение обнаруженных свойств генетических молекул ДНК и РНК. Исследование включает три этапа: параметризация, масштабирование и визуализация. Параметризация — определение учитываемых параметров, которые основаны на структурных и физико-химических свойствах нуклеотидов как элементарных составных частей генома. Масштабирование играет роль «фокусировки» и позволяет исследовать генетические структуры в различных масштабах. Визуализация включает выбор осей координатной системы и способа визуального отображения. Представленные в работе алгоритмы выдвигаются на роль расширенного инструментария для развития научно-исследовательского программного обеспечения анализа длинных нуклеотидных последовательностей с возможностью отображения геномов в параметрических пространствах различной размерности. Одним из значимых результатов исследования является то, что были получены новые биологически интерпретируемые критерии классификации геномов различных живых организмов для выявления межвидовых взаимосвязей. Новая концепция позволяет визуально и численно оценить вариативность физико-химических параметров нуклеотидных последовательностей. Эта концепция также позволяет обосновать связь параметров молекул ДНК и РНК с фрактальными геометрическими мозаиками, обнаруживает упорядоченность и симметрии полинуклеотидов и их помехоустойчивость. Полученные результаты стали обоснованием для введения новых научных терминов: «генометрия» как методология вычислительных стратегий и «генометрика» как конкретные параметры того или иного генома или нуклеотидной последовательности. В связи с результатами исследования затронуты вопросы биосемиотики и уровни иерархичности организации живой материи.

Ключевые слова: генетические алгоритмы, вариативность, многомерный анализ данных, физико-химические параметры нуклеиновых кислот, конечная геометрия

**COMPUTER RESEARCH AND MODELING**
**2020 VOL. 12 NO. 2 P. 417–434**
DOI: 10.20537/2076-7633-2020-12-2-417-434

Ки&М

**ANALYSIS AND MODELING OF COMPLEX LIVING SYSTEMS**

UDC: 519.688, 51-76

# Biomathematical system of the nucleic acids description

**I. V. Stepanyan**

Mechanical Engineering Research Institute of the Russian Academy of Sciences (IMASH RAN),
4 Maly Kharitonyevsky Pereulok, Moscow, 101990, Russia

E-mail: neurocomp.pro@gmail.com

*Received 11.10.2019, after completion — 17.12.2019.*
*Accepted for publication 26.12.2019.*

The article is devoted to the application of various methods of mathematical analysis, search for patterns and studying the composition of nucleotides in DNA sequences at the genomic level. New methods of mathematical biology that made it possible to detect and visualize the hidden ordering of genetic nucleotide sequences located in the chromosomes of cells of living organisms described. The research was based on the work on algebraic biology of the doctor of physical and mathematical sciences S. V. Petukhov, who first introduced and justified new algebras and hypercomplex numerical systems describing genetic phenomena. This paper describes a new phase in the development of matrix methods in genetics for studying the properties of nucleotide sequences (and their physicochemical parameters), built on the principles of finite geometry. The aim of the study is to demonstrate the capabilities of new algorithms and discuss the discovered properties of genetic DNA and RNA molecules. The study includes three stages: parameterization, scaling, and visualization. Parametrization is the determination of the parameters taken into account, which are based on the structural and physicochemical properties of nucleotides as elementary components of the genome. Scaling plays the role of "focusing" and allows you to explore genetic structures at various scales. Visualization includes the selection of the axes of the coordinate system and the method of visual display. The algorithms presented in this work are put forward as a new toolkit for the development of research software for the analysis of long nucleotide sequences with the ability to display genomes in parametric spaces of various dimensions. One of the significant results of the study is that new criteria were obtained for the classification of the genomes of various living organisms to identify interspecific relationships. The new concept allows visually and numerically assessing the variability of the physicochemical parameters of nucleotide sequences. This concept also allows one to substantiate the relationship between the parameters of DNA and RNA molecules with fractal geometric mosaics, reveals the ordering and symmetry of polynucleotides, as well as their noise immunity. The results obtained justified the introduction of new terms: "genometry" as a methodology of computational strategies and "genometrica" as specific parameters of a particular genome or nucleotide sequence. In connection with the results obtained, biosemiotics and hierarchical levels of organization of living matter are raised.

Keywords: genetic algorithms, variability, multivariate data analysis, chemical parameters of nucleic acids, finite geometry

Citation: *Computer Research and Modeling*, 2020, vol. 12, no. 2, pp. 417–434 (Russian).

# 1. Введение

Одной из основных задач математической биологии является развитие численных методов, которые позволяют получать новые знания и создавать алгоритмы для исследования биологических объектов и явлений. Существует проблема восприятия сложной биологической информации, в частности процессов, происходящих внутри клетки. Эта проблема относится к психофизиологии восприятия любой сложной и многомерной информации. Дело в том, что наглядно визуализировать реальные процессы, происходящие в живой клетке, довольно затруднительно, несмотря на серьезную теоретическую базу, наличие проработанного математического аппарата и возможности современной компьютерной графики. Это же относится и к задаче анализа вариативности физико-химических параметров длинных нуклеотидных последовательностей (нуклеиновых кислот) ДНК и РНК. Таким образом, решение проблемы визуализации сложных биологических данных позволяет усовершенствовать научно-методологическую базу исследований за счет появления новых и удобных инструментов, улучшающих восприятие генетической информации.

Среди программных продуктов (ПО) для анализа генетических нуклеотидных последовательностей (имеются в виду данные вида AGGCT…, полученные из ДНК живых организмов и хранящиеся в файлах или базах данных) в основном представлены алгоритмы, основанные на статистическом анализе. Однако эти инструменты не обладают широкими визуальными возможностями, поскольку результаты их выполнения (графики, гистограммы, таблицы и пр.) не достаточно выразительны (понятны исключительно специалистам), что сужает круг пользователей соответствующего ПО. Также для анализа сложных многопараметрических явлений, к которым относится исследуемый в данной работе феномен генетического кодирования, применяются методы понижения размерности и машинное обучение. Существуют методы компьютерной графики, которые позволяют построить некоторое упрощение исследуемых объектов в виде компьютерных моделей. Вклад в развитие когнитивной графики сделал А. А. Зенкин [Зенкин, 1991], который исследовал алгоритмы двумерного представления одномерных процессов. Также известна пионерская работа Р. М. Акселрода [Axelrod, 1976] по визуализации информации. В настоящее время данное направление интенсивно развивается и представлено в [Eidenzon, Pilipczuk, 2015]. При этом применительно к генетической информации когнитивная графика принципиально не продвинулась дальше, чем схематическое представление спирали ДНК и внутриклеточных процессов.

Строгие алгебраические подходы к исследованию биологических явлений обосновал и ввел в математическую биологию С. В. Петухов в работах [Петухов, 2008, 2010–2012, 2017], продолжающих исследования биофизика Румера [Румер, 1968]. В частности, С. В. Петуховым были обнаружены и исследованы многомерные поличисловые алгебры, описывающие матричные принципы феноменологии генетического кодирования как системы резонансов в тензорных семействах матриц [Петухов, 2012]. Им было введено понятие таблиц наследования собственных значений матриц и показана их аналогия с решетками Пеннета полигибридного скрещивания организмов по законам Менделя. Он показал, что аллели генов можно интерпретировать как резонансы (собственные значения матриц) некоторых колебательных систем. Им также были обнаружены новые для математического естествознания числовые системы, развивающие теорию связи и помехоустойчивого кодирования [Петухов, 2008; Petoukhov, He, 2010]. Характерно, что исследования в области алгебраической биологии носят междисциплинарный характер.

В данной работе описана новая фаза развития матричных методов исследования свойств нуклеотидных последовательностей (и их физико-химических параметров), построенная на принципах конечной геометрии. Вопросами исследования знаковых систем в живой природе занимается биосемиотика, в контексте которой рассматриваются семиотика порождения текста и семиотика рецепции (восприятия) текста [Седов, 2000]. Алгоритмы когнитивной компьютерной графики, примеры которой представлены в настоящей работе, относятся к семиотике вос-

приятия генетических текстов и позволяют биологически интерпретировать те свойства нуклеотидных последовательностей, которые сложно воспринять путем их простого прочтения или стандартными средствами статистического анализа.

Автором предложено введение новых терминов: «генетическая геометрия (геометрия)» и «геометрика». Геометрия — научное направление в области молекулярно-биологической семиотики, основанное на применении методов конечной геометрии и представляющее собой набор подходов к исследованию внутриструктурных физико-химических свойств генома в различных параметрических координатных системах и на различных масштабах визуализации. Геометрика — индивидуальная визуализация физико-химических параметров той или иной конкретной генетической нуклеотидной последовательности (ДНК или РНК) либо ее фрагмента.

## 2. Постановка задачи. Генетический код как система субалфавитов

Нуклеиновые кислоты ДНК и РНК — дискретные последовательности нуклеотидов, выполняющие функции хранения, обработки и передачи наследственной генетической информации в живых организмах [Chargaff, 1952; Crick, 1979]. Такие последовательности анализируются, как правило, статистическими методами. Они имеют одномерный линейный характер и хранятся на компьютерах в виде строк, состоящих из букв фиксированного алфавита, кодирующего нуклеотиды: цитозин (C), тимин (T), урацил (U), аденин (A), гуанин (G). Тимин заменен урацилом при переходе из ДНК в РНК (рис. 1 и 2).

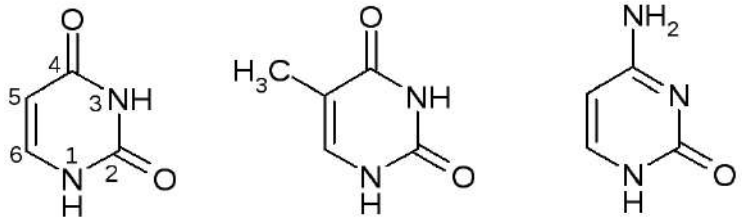


Рис. 1. Пиримидиновые основания (слева направо): урацил (2,4-диоксопиримидин), тимин (5-метил-2,4-диоксопиримидин), цитозин (2-оксо-4-амино-пиримидин)

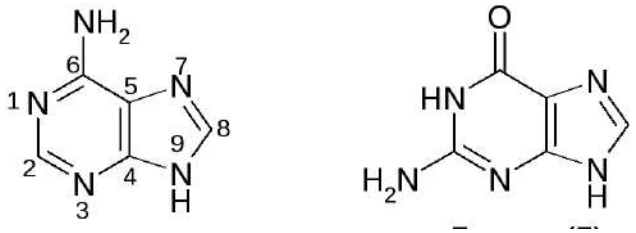


Рис. 2. Пуриновые основания: аденин (6-аминопурин) (слева), гуанин (2-амино-6-оксопурин) (справа)

Химические формулы пуриновых и пиримидиновых оснований приведены для демонстрации бинарно-оппозиционных свойств их физико-химических параметров: каждое азотистое основание генетического кода имеет три варианта своего двоичного представления. Эти варианты представлений, названные бинарными субалфавитами, различаются в соответствии с типами бинарно-оппозиционных свойств в наборе азотистых оснований:

1) G = C «3 водородные связи» / A = T «2 водородные связи»;
2) C = T «пиримидины» / A = G «пурины»;
3) A = C «амино» / G = T «кето».

Таким образом, структура химических формул составляет систему базисных ортогональных функций Уолша [Balonin et al., 2017], «прошитой» в любой молекуле ДНК или РНК в соот-

ветствии со структурой и свойствами составляющих ее нуклеотидов. С учетом дополнительного (нулевого) признака (наличие фосфатного остатка), который не является оппозиционным, система генетических субалфавитов может быть представлена в виде матрицы Адамара на рис. 3. Данная матрица представляет собой набор ортогональных Уолш-функций, является симметричной, поскольку нуклеотиды могут быть заменены соответствующими субалфавитами без изменения структуры матрицы.

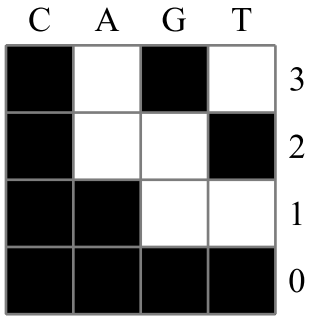



Рис. 3. Вариант матрицы Адамара, отображающей кодирование нуклеотидных субалфавитов. Затемненные клетки +1, белые клетки –1 (или, наоборот, в зависимости от способа кодирования). Номера субалфавитов обозначены как 1, 2, 3 и 0

Казалось бы, что сама гипотеза очевидна, поскольку любые символьные последовательности могут быть оцифрованы, преобразованы и визуализированы, причем бесконечным числом способов. Однако в рассматриваем в данной статье методе число способов кодирования нуклеотидов по указанным физико-химическим признакам конечно и влияет на трансформации симметрии в итоговой визуализации. Число таких вариантов кодирования совпадает с числом трансформаций симметрии итоговых паттернов и зависит от способа кодирования, т. е. от вида матрицы Адамара на рис. 3 (что закрашивать черным, что белым цветом). Любое нарушение кодирования бинарно-оппозиционных признаков влечет построение неинформативных визуализаций. Таким образом, это вариант алгебраического кодирования, основой которого является взаимосвязанный набор алгебраических операций [Сагалович, 2011].

Последовательности ДНК и РНК являются фундаментальными для кодирования и обработки генетической информации. Генетическая информация может быть интерпретирована не только из символьных последовательностей, но также из скрытых сигналов внутри самих последовательностей [Yin, 2019]. Гипотеза исследования состоит в том, что символьные последовательности могут быть преобразованы методами цифровой обработки сигналов в числовые последовательности и затем визуализированы в пространствах двоично-ортогональных функций Уолша, чтобы скрытые сигналы могли быть выявлены.

## 3. Семантический метод визуализации нуклеотидных последовательностей на примере гена инсулина

Предлагаемый семантический подход к визуализации представляет собой методологию, заключающуюся в построении такого графа, вершинами которого являются фрагменты генетической последовательности равной длины $N$ («слова» или семантические единицы), а дугами — связи, которые отображают соседство фрагментов. Дуги могут быть направленными, отображая соседство справа или слева. Длина фрагментов является свободным параметром алгоритма, что позволяет построить для одной и той же нуклеотидной последовательности графы в различных масштабах. Примеры построения семантической сети гена инсулина при разбиении генетической последовательности нуклеотидов на фрагменты различной длины приведены на рис. 4–6.

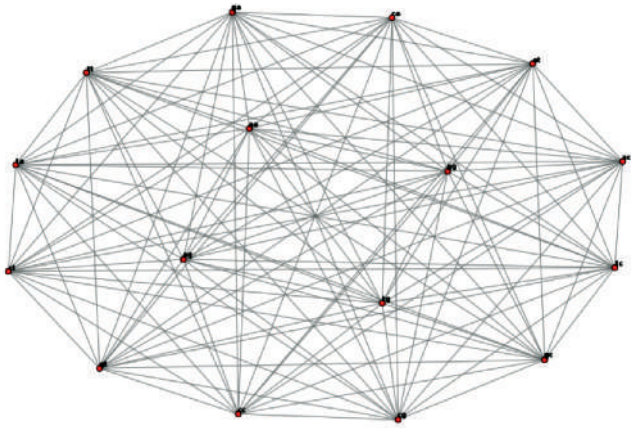


Рис. 4. Семантическая сеть гена инсулина при разбиении кодирующей нуклеотидной последовательности на блоки длиной 2 нуклеотида

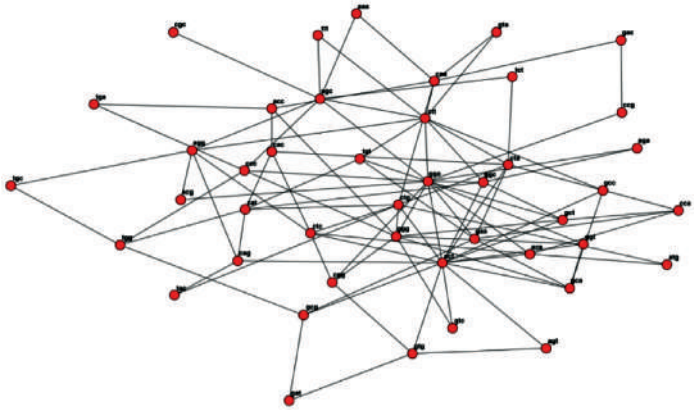


Рис. 5. Семантическая сеть гена инсулина при разбиении кодирующей нуклеотидной последовательности на блоки длиной 3 нуклеотида

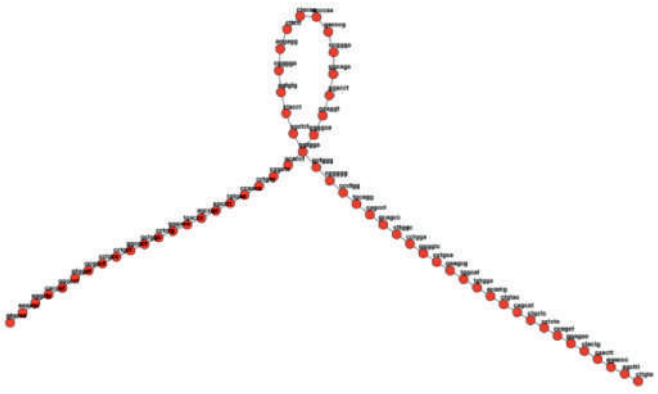


Рис. 6. Семантическая сеть гена инсулина при разбиении кодирующей нуклеотидной последовательности на блоки длиной 6 нуклеотидов. Видна функциональная роль «узлового» 6-плета

Из графов, отображающих семантические связи одного и того же гена (рис. 4–6), видно, что можно оценить масштабную вариативность его структуры в зависимости от параметра масштаба. При этом на различных масштабах структура графа сильно меняется. Видно, что с увеличением длины слова связность графа уменьшается, и можно выделить такие узловые элементы, как старт- и стоп-кодоны. Однако структура графа далеко не всегда информативна.

Для больших геномов, которые встречаются в живой природе, структура семантического графа может напоминать запутанный клубок, что сильно усложнит восприятие генетической информации. Можно анализировать соответствующие графам матрицы связности, но отсутствие выраженных симметрий, отражающих нуклеотидный состав, не дает принципиально нового качества.

Семантические графы приведены как пример подхода к визуализации генетической информации, который недостаточно удобен для анализа. В частности, структуры графов сложны для восприятия и не позволяют визуализировать скрытые симметрии и закономерности нуклеотидного состава. Причем очевидно, что субалфавитный уровень анализа семантических графов также возможен, но теоретико-графовое (или «клубочное») представление в данном случае рассматривается как менее эргономичное. В связи с этим ставится задача разработки матричных методов и представлений на основе бинарных субалфавитов. Для более информативной оценки использовали соображения, разработанные и изложенные в [Stepanyan, Petoukhov, 2017]. Эти соображения основаны с учетом:

- метода семантической визуализации (наличие свободного параметра масштабирования);
- структуры физико-химических параметров нуклеотидов (бинарных субалфавитов), приведенных выше.

## 4. Алгоритм масштабно-параметрического моделирования физико-химических параметров длинных нуклеотидных последовательностей

Моделируются физико-химические параметры последовательностей по трем бинарным субалфавитам, закодированным в матрице Адамара на рис. 3. Следует отметить, что для стандартизации исследований целесообразно договориться о единственном представлении данной матрицы как способа кодирования, с тем чтобы полученные по представленному алгоритму генометрические визуализации совпадали между собой. В противном случае (при различном способе кодирования признаков черным или белым цветом в ячейках матрицы) можно наблюдать симметрические отображения итоговых визуализаций.

Приведенный авторский алгоритм лежит в основе построения масштабно-параметрической модели генома для визуализации в координатных пространствах различной мерности и топологии.

1. Масштабирование. Последовательность символов из набора {A, G, C, T} или {A, G, C, U}, кодирующих азотистые основания, разделяется на фрагменты равной длины $N$, где $N$ — свободный параметр алгоритма — длина «слова». Полученные фрагменты равной длины назовем $N$-мерами или $N$-плетами.

2. Параметризация. С учетом системы генетических субалфавитов (рис. 3) последовательность азотистых оснований может быть представлена в виде трех бинарных последовательностей, состоящих из нулей и единиц. Выбор способа кодирования (что считать нулем или единицей) влияет на преобразования симметрии итоговой визуализации.

3. Визуализация. Полученная бинарная запись фрагментов является их представлением в виде трех последовательностей десятичных или иных однозначно идентифицирующих значений. Преобразование двоичных $N$-плетов в десятичные числа позволяет отобразить их в выбранной системе координат. Полученные значения задают координаты точек в пространстве параметров (далее — в пространстве визуализации или параметрическом пространстве).

Шаги 1 и 2 могут быть переставлены (сначала параметризация, затем масштабирование), что может повлиять на вычислительную нагрузку при обсчете длинных последовательностей, так как для оптимизации и распараллеливания вычислений целесообразна одна нарезка исходной последовательности вместо трех по каждому из субалфавитов. В связи с большим счетным временем полученные десятичные результаты нарезки были сохранены в промежу-

точном json-файле для последующей визуализации различными графическими библиотеками. При анализе ДНК моделирование идет по одной цепи — комплементарная сеть порождает идентичную картинку, повернутую относительно первоначальной, поэтому не рассматривается.

С учетом того, что этап параметризации может быть привязан к различным параметрам генетического кодирования, алгоритм можно рассматривать в качестве методологической базы системы биоматематических методов моделирования в молекулярной генетике.

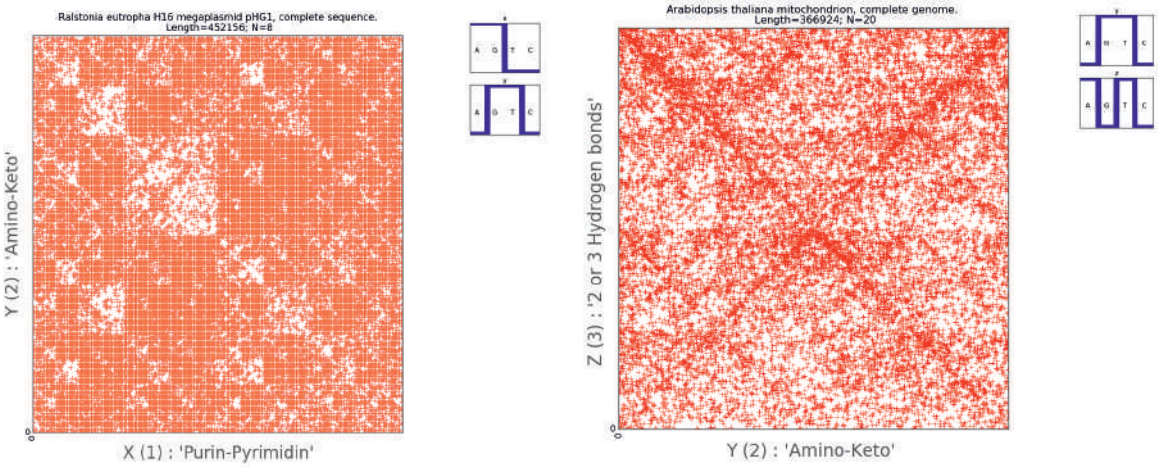
В результате применения алгоритма задается модельное пространство, которое является параметрическим, конечным и дискретным. Комбинаторные свойства этого пространства позволяют отобразить любые полинуклеотиды для произвольного конечного $N$. Упорядоченные числовые значения на координатных осях отображают физико-химические характеристики $N$-меров, поскольку они однозначно задаются свойствами бинарно-оппозиционных субалфавитов. Отметим, что общенаучные методы изучения нуклеиновых кислот обычно концентрируют свое внимание на те фрагменты, которые в них присутствуют. Предложенные алгоритмы позволяют представить в наглядной форме феноменологию и особенности дефицита и присутствия различных типов $N$-меров. Алгоритм масштабно-параметрического моделирования был применен для анализа различных молекул РНК и ДНК. В ходе исследований было визуализировано около сотни геномов из базы данных NCBI (простейшие, растения, грибы, животные, вирусы). Для демонстрации результатов визуализаций в данной статье использовались некоторые отобранные последовательности, которые по результатам исследований продемонстрировали выраженный индивидуальный характер строения геометрической мозаики. Часть расчетов была выполнена на суперкомпьютере «МВС-10П» (МСЦ РАН).

## 5. Примеры визуализации вариативности физико-химических параметров нуклеотидных последовательностей

Набор трех бинарно-оппозиционных признаков может быть сопоставлен с осями $\{X, Y, Z\}$ декартовой системы координат. Полученные таким образом трехмерные представления нуклеиновых кислот не достаточно удобны для восприятия и анализа их особенностей. Но двумерные проекции трехмерных представлений подходят для отображения специфики строения генетических последовательностей. В базисах $\{X, Y\}$, $(X, Z\}$ и $\{Y, Z\}$, выбранных в качестве декартовых систем координат, получаем три различные двумерные проекции на основе пар соответствующих субалфавитов физико-химических параметров нуклеотидов (субалфавиты обозначены на рис. 7, *а*, *б* справа вверху). Исходя из свойства матрицы Адамара (рис. 3), согласно которому тройка бинарно-оппозиционных субалфавитов связана между собой операцией сложения по модулю два, для определения произвольной нуклеиновой кислоты достаточно любой пары ее бинарных представлений. Поэтому для однозначно интерпретируемой двумерной визуализации нуклеотидного состава достаточно любой пары двумерных проекций из трех.

На рис. 7 приведены примеры двумерной визуализации геномов различных организмов, рядом в порядке A, G, T, C приведены пары функций Уолша, которые использовались для кодирования признаков. Используемый метод представлен в § 4 («Алгоритм масштабно-параметрического моделирования физико-химических параметров длинных нуклеотидных последовательностей»). Анализ двумерных отображений, полученных на основе предложенного алгоритма визуализации, позволяет выявить особенности нуклеотидного состава у хромосом различных видов организмов. При этом ДНК некоторых видов организмов обнаруживают некоторый «шаблонный» нуклеотидный состав.

О биологическом значении метода свидетельствует тот факт, что генерированные случайным образом последовательности при алгоритмической визуализации дают паттерн, все точки которого разбросаны хаотично. Нами были сгенерированы случайным образом последовательности азотистых оснований длиной в 100000 нуклеотидов с разбиениями на $N$-плеты различной длины. Случайные визуальные представления носят нерегулярный, хаотический характер, при полном отсутствии каких-либо мозаик и симметрий по всем субалфавитам, что значительно




(а) (б)

Рис. 7. Двумерная генометрика нуклеотидного состава генома бактерии *Ralstonia eutropha* при параметре масштабирования $N = 8$ (а). Иллюстрация двумерного представления нуклеотидного состава генома митохондрии растения Резуховидка (резушка) Таля (лат. *Arabidopsis thaliana*) семейства капустные (*Brassicaceae*) при $N = 20$ (б). Осям абсцисс и ординат соответствуют десятичные представления двоичного кодирования каждого $N$-плета

отличает их от реальных длинных нуклеотидных последовательностей, находящихся в молекулах ДНК. Таким образом, предложенные алгоритмы генометрической визуализации представляются полезными для исследования скрытых закономерностей в хромосомах живых организмов, а также для возможности визуальной классификации и сравнительного анализа различных геномов.

Масштабная инвариантность — свойство сохранения при изменении всех расстояний в одинаковое число раз. Такие преобразования подобия образуют группу масштабных преобразований. В задаче анализа нуклеотидного состава масштабная вариативность представляет собой устойчивость визуального паттерна к масштабированию (изменению параметра $N$) и выражена у различных организмов в различной степени. Пример визуализации генома при различных параметрах масштаба $N$ приведен на рис. 8.

Из рис. 8 видно, что изменение масштабирующего параметра $N$ позволяет исследовать геном на различных уровнях детализации. Этот параметр позволяет регулировать фокусировку изображения. При определенном $N$ картинка становится четкой, мозаика приобретает выраженный характер и, как правило, начинают прослеживаться фрактальные паттерны. Таким образом, масштабирующий коэффициент $N$ играет роль разрешающей способности геометрической визуализации: большие $N$ дают малое число точек, малые $N$ дают малую координатную сетку. Это обстоятельство позволяет говорить о разномасштабном анализе в многомерных параметрических пространствах с применением алгебраического кодирования.

Результаты применения алгоритма двумерной визуализации позволяют сделать вывод о высокой стабильности итоговых мозаик при зашумлениях исходной последовательности, в том числе при сдвигах рамки считывания последовательностей, в случаях удаления произвольных фрагментов последовательности (проряжениях), при реверсировании всей анализируемой цепи или ее фрагментов, при различных типах перестановок $N$-меров и нуклеотидов (в ряде случаев вплоть до полной перестановки всех нуклеотидов в последовательности). В частности, наблюдали стабильность мозаичных узоров в случаях удаления каждого второго нуклеотида, каждого третьего нуклеотида и т. д. С учетом того, что, по результатам экспериментов, зашумления в длинных нуклеотидных последовательностях существенно не влияют на форму итогового паттерна, все имеющиеся пропуски в длинных последовательностях алгоритмически игнорировались. При этом визуализация нуклеиновых кислот в двумерных пространствах

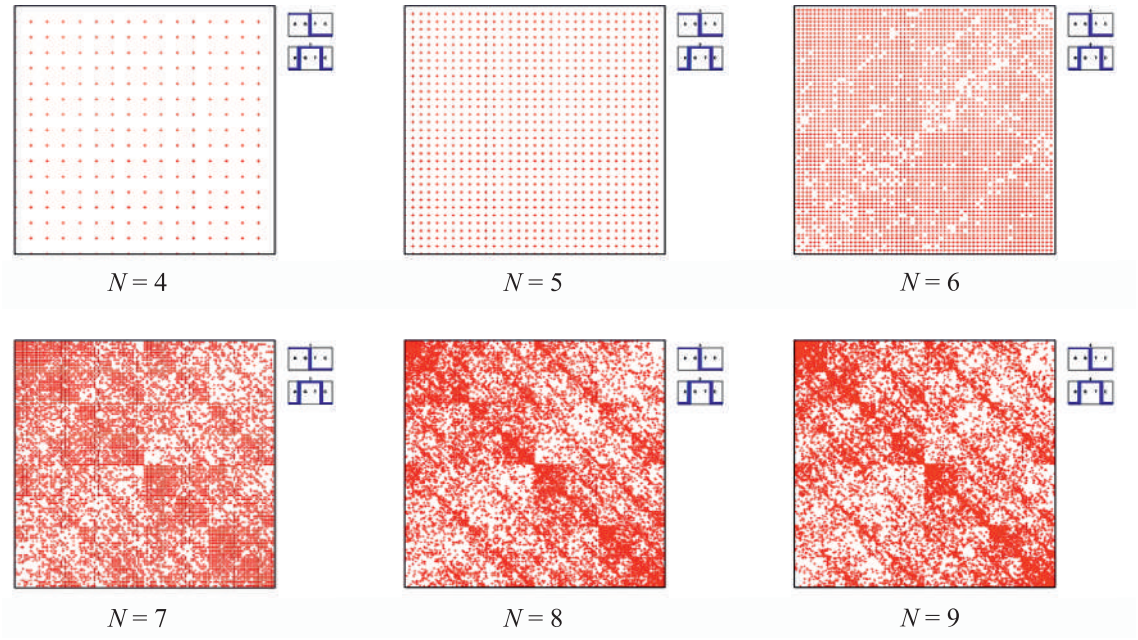



Рис. 8. Двумерная геометрика нуклеотидного состава генома хлоропласты Fistulifera sp. JPCC DA0580 при различных параметрах масштабирования $N$ от 4 до 9. Осям абсцисс и ординат соответствуют десятичные представления двоичного кодирования каждого $N$-плета. В соответствии с Уолш-функциями (приведенными рядом с каждой визуализацией) ось абсцисс кодирует признаки 2-го субалфавита «пиримидины» / «пурины», ось ординат — признаки 3-го суб-алфавита «амино» / «кето»

в ряде случаев характеризуется выраженными симметриями и стабильностью не только к зашумлениям в исходных данных, но и к различным значениям параметра масштаба $N$ в пределах некоторого диапазона. Это демонстрирует тотальную надежность и помехоустойчивость в генетических системах.

Отметим, что двумерные модели, построенные с помощью предложенного алгоритма визуализации, напоминают фрактальные паттерны нуклеиновых кислот, которые были получены с помощью метода CGR [Feldman, David, 2012]. Метод CGR известен и исследуется как технология картирования последовательности генома, которая преобразует последовательности генома в двумерные изображения наподобие рис. 7. Исследование [Pei et al., 2019] дает существенное понимание изучения филогении, эволюции и эффективных алгоритмов сравнения ДНК для больших геномов на основе данного метода. В работе [Duarte-sanchez et al., 2018] мультифрактальный анализ с использованием метода CGR позволил количественно оценить генетическую изменчивость и нелинейную стабильность последовательности генома человека для объяснения некоторых генетических заболеваний, вызванных аномалиями хромосом; применение фрактального анализа используется в исследованиях по генетике рака для исследования хаотического поведения мутаций и разработки новых методов лечения [Hewelt et al., 2019]. Также изложенные в данной работе результаты перспективны для развития такого направления, как ДНК-оригами [Tikhomirov et al., 2017], для обеспечения формирования нанометровых паттернов, которые можно использовать для создания программируемых молекулярных машин и массивов функциональных материалов, конструирования сложных материалов и устройств с размерами, аналогичными размерам бактерии.

Однако в методе CGR не учитывается информация о бинарных субалфавитах, вследствие чего данный метод воспринимается как некий алгоритмический феномен, т. е. не интерпретируется с позиции бинарных субалфавитов физико-химических признаков. Тем не менее геометрический подход обнаруживает связь с системами итерированных функций Дж. Хатчинсона [Hutchinson, 1981] и М. Барнсли [Barnsley, 1988], которую можно рассматривать как вариант построения теоретической базы для математического исследования генетических феноменов.

Для дальнейших рассуждений еще раз отметим, что бинарные субалфавиты связаны между собой операцией сложения по модулю «два» и тем самым задают модельное пространство со свойствами, при которых координаты всех точек являются как бы «склеенными» этой операцией. Это следует из свойств матрицы Адамара на рис. 3. В связи с этим имеет смысл рассматривать каждый субалфавит в отдельности как отдельное измерение, в котором ДНК параметризуется по всей длине. При геометрической визуализации ось абсцисс кодирует порядковый номер *N*-плета в генетической последовательности, а ось ординат отображает десятичные значения двоичного представления каждого *N*-плета. Соответствующую визуализацию можно считать параметрически одномерной в силу способа отображения молекулы.

Таким образом, существует три одномерные визуализации, между которыми действительна алгебраическая связь. Использование одномерных координатных осей $\{X\}$, $\{Y\}$ и $\{Z\}$ дает тройку отображений с использованием соответствующих субалфавитов.

Отметим, что особенностью геометрии является возможность сжатия. Как видно, последовательность ДНК может быть кодирована и восстановлена с использованием трех целых десятичных чисел: одно целое число представляет собой длину последовательности, а два других целых числа представляют два параметризованных бинарных субалфавита (рис. 3).

На рис. 9 приведен пример визуализации хромосомы человека, где хорошо видны области с различным нуклеотидным составом. Эти специфические регионы могут быть визуализированы в двумерных параметрических пространствах для их дальнейшего исследования. Используемый метод представлен в § 4 («Алгоритм масштабно-параметрического моделирования физико-химических параметров длинных нуклеотидных последовательностей») с той особенностью, что на шаге 3 две оси служат для отображения физико-химических параметров, а одна — для отображения порядкового номера *n*-плета в последовательности (пространственная ось). Поскольку учитывали три субалфавита, то получили три возможных варианта пар признаков, каждая из которых сопоставлялась со своей пространственной осью.

Области хромосомы с различными физико-химическими параметрами имеют индивидуальный характер визуализации в силу особенностей различий в нуклеотидном составе. Эти области хорошо видны на рис. 9. Данный вариант одномерной параметрической визуализации позволяет отобразить состав молекулы в ее пространственном расположении, что в некотором смысле ближе к физическому пространству, чем к двумерному параметрическому, поскольку молекула ДНК в данном случае представлена по каждому из своих субалфавитов по всей своей протяженности. При этом получили частичную аналогию с физическим пространством, поскольку, как было отмечено выше, параметрическое пространство задается так, что ось абсцисс используется для отображения порядкового номера *n*-плета в последовательности (в то время как ось ординат — для визуализации физико-химических параметров). В связи с этим алгоритмы одномерной визуализации оказываются информативными для оценки изменений физико-химических параметров молекулярных последовательностей ДНК и РНК с возможностью привязки к конкретным фрагментам этих молекул и возможностью их многомасштабного анализа.

Обобщая вышесказанное, отметим, что любую хромосому или другую генетическую нуклеотидную последовательность можно представить как набор двумерных паттернов, следующих друг за другом. Для этого необходима нарезка последовательности на блоки, каждый из которых анализируется в двумерном пространстве по отдельности. Полученные двумерные паттерны при визуализации выстраиваются друг за другом. В связи с этим возможен еще один алгоритм визуализации:

- строится три одномерных представления по каждому из субалфавитов (как на рис. 9);
- области, где присутствуют изменения в нуклеотидном составе, анализируются двумерными представлениями.

Еще один подход в геометрии основан на подсчете общего числа пуринов, пиримидинов и других химических признаков в каждом *N*-мере. Используемый метод представлен в § 4 («Алгоритм масштабно-параметрического моделирования физико-химических параметров длинных нуклеотидных последовательностей») с той особенностью, что на шаге 2 в соответствии с методами теории секвентного анализа Хармута [Хармут, 2016] строились визуализации

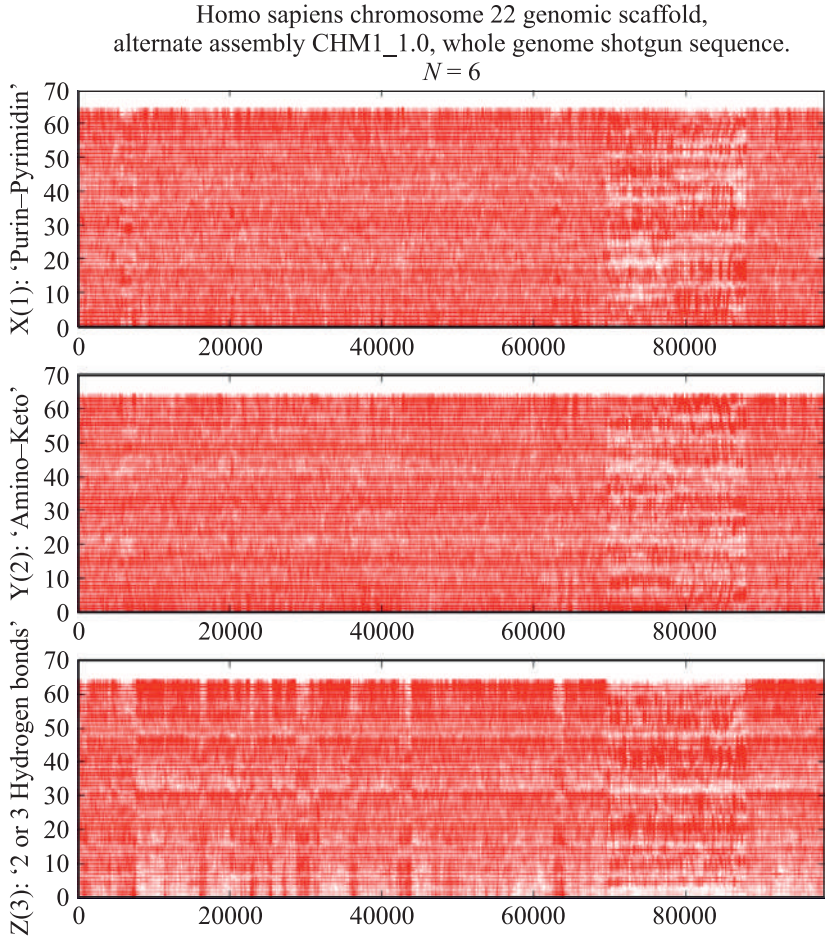



Рис. 9. Визуализация трехканального представления нуклеотидного состава фрагмента первой хромосомы человека. Каждому из трех рядов соответствует двоично-оппозиционный субалфавит. Ось абсцисс кодирует порядковый номер $N$-плета, ось ординат — бинарный код $N$-плета по соответствующему субалфавиту

по количеству элементов (нулей или единиц), которые встречались в двоичных представлениях $N$-плетов в последовательностях азотистых оснований. В связи с тем, что этот способ основан на суммарном числе тех или иных параметров, соответствующие пространства визуализации будем называть интегральными. На рис. 10 приведен пример интегрально-параметрического представления нуклеотидного состава хромосомы человека в одной из плоскостей двумерной визуализации.

Интегральные двумерные визуализации реальных генетических последовательностей имеют вид пятна, как правило вытянутого по горизонтали или по вертикали. Они менее информативны, чем соответствующие им двумерные фрактальные мозаики. Однако они удобны при визуализациях наподобие рис. 11.

На рис. 11 показан пример визуализации трехканального представления нуклеотидного состава фрагмента 22-й хромосомы человека. Используемый метод представлен в § 4 («Алгоритм масштабно-параметрического моделирования физико-химических параметров длинных нуклеотидных последовательностей») с той особенностью, что на шаге 2 в соответствии с методами теории секвентного анализа Хармута [Хармут, 2016] строились визуализации по количеству элементов (нулей или единиц), которые встречались в двоичных представлениях $N$-плетов в последовательностях азотистых оснований, а на шаге 3 две оси служат для отображения физико-химических параметров, а одна — для отображения порядкового номера $n$-плета в последовательности. Поскольку учитывали три субалфавита, то получили три возможных варианта пар признаков, каждая из которых сопоставляется своей пространственной оси.

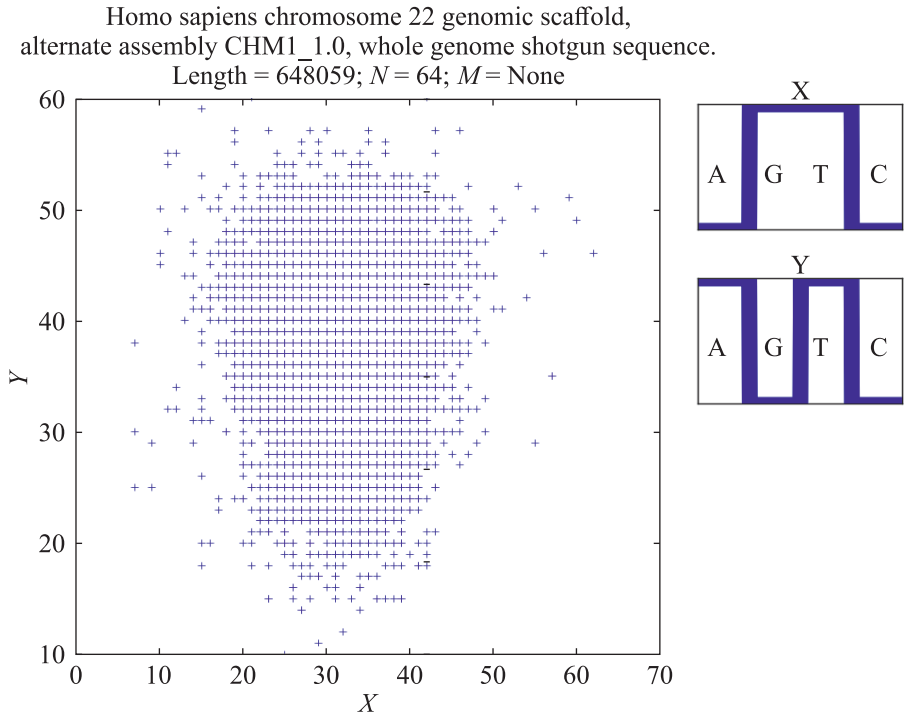



Рис. 10. Иллюстрация интегрально-двумерного представления нуклеотидного состава хромосомы человека по одной из плоскостей визуализации. Пара Уолш-функций, используемых для параметризации отображена справа. Осям абсцисс и ординат соответствует количество единиц каждого 64-мера с использованием пары двоично-оппозиционных субалфавитов: ось абсцисс соответствует третьему бинарному субалфавиту «амино»/«кето», ось ординат соответствует первому бинарному субалфавиту «3 водородные связи» / «2 водородные связи»

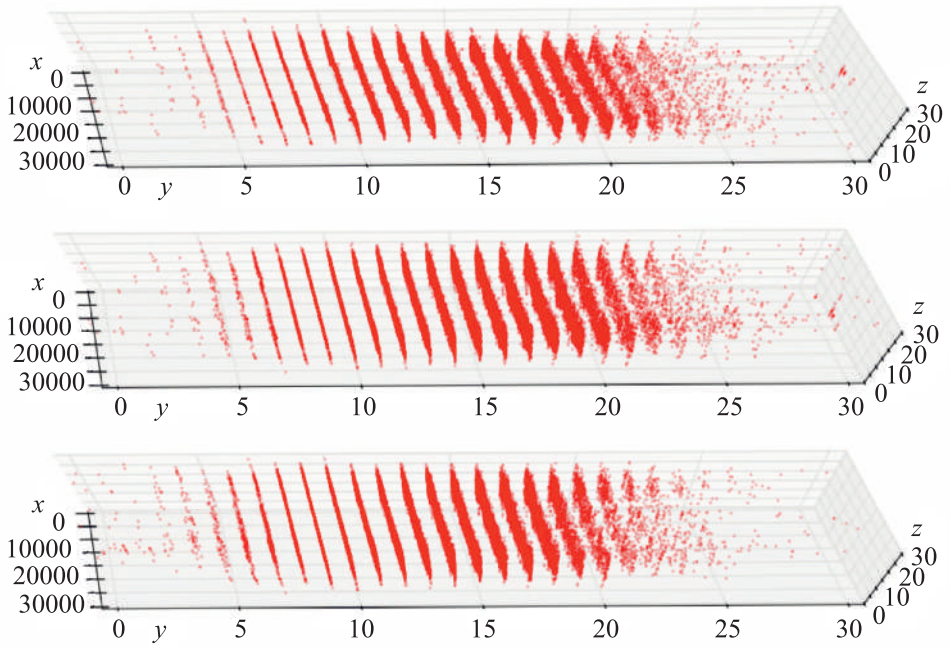



Рис. 11. Геометрика трехканального представления нуклеотидного состава фрагмента 22-й хромосомы Homo Sapiens. Отображена интегральная визуализация суммарного числа единиц в кодах *N*-меров по каждой из трех пар субалфавитов с нарезкой хромосомы на «окна» равной длины. Ось *x* отображает порядковый номер мозаики двумерной геометрической визуализации. Каждой из трех проекций соответствует пара двоично-оппозиционных субалфавитов, построенных по осям *y* и *z*

Это вариант объемного изображения, которое позволяет визуально оценить физико-химические характеристики хромосомы. Предварительно была произведена дополнительная нарезка на окна двумерной визуализации. Каждое двумерное окно располагается за другим и представляет собой интегральное представление как на рис. 10. Этот вариант визуализации полезен для оценки изменений в нуклеотидном составе при прочтении фрагмента молекулы от начала до конца. Глубина регистрируемых изменений определяется масштабирующим параметром *N* алгоритма и шагом нарезки. Предлагается новый инструмент, который, по мнению автора, содержит в себе потенциал для дальнейшего использования другими исследователями.

При интегральных визуализациях сохраняется параметр масштабирования, позволяющий регулировать информативность результата. В связи с тем, что на различных участках молекулы находится различный нуклеотидный состав, рациональная алгоритмическая подстройка параметра масштабирования является отдельной задачей и требует дополнительных исследований при решении практических вопросов в рамках изложенного подхода.

В целом изложенные идеи по представлению генетической информации в различных параметрических пространствах открывают новые возможности для упрощения изучения геномов с применением различных метрик и топологий (цилиндрическое, сферическое и другие пространства), поскольку физико-химические параметры нуклеиновых кислот могут быть привязаны к любым геометриям на основе изложенного алгоритма. При этом сохраняется алгоритмическая связь с реальными данными, что делает метод биологически интерпретируемым благодаря матрице Адамара (рис. 3), по которой кодируются и алгоритмически строятся все визуальные отображения.

## 6. Использование алфавитов 16 дуплетов

В [Petoukhov, 2017] приведены субалфавиты, которые были получены при анализе мозаичной матрицы дуплетов $[C\ A;\ T\ G]^{(2)}$ как варианты декомпозиции этой матрицы (в данном случае кронекеровская степень матрицы отвечает за все варианты пар нуклеотидов). Эти субалфавиты приведены в таблице 1. В результате применения алгоритма визуализации мы получили паттерны, примеры которых приведены на рис. 12. В данном варианте представления каждой точке соответствует последовательность дуплетов, так как масштабирующий параметр *N* характеризует количество пар нуклеотидов в каждой точке.

Таблица 1. Субалфавиты для систем 2-блочных U-комплексных чисел, которые С. В. Петухов получил при анализе Уолш-представления мозаичной геноматрицы $[C\ A;\ T\ G]^{(2)}$

Первый субалфавит UC1:

- $CC = CG = TC = TG = 00_e$
- $TT = TA = CA = CT = 01_e$
- $GG = GC = AG = AC = 10_e$
- $AA = AT = GT = GA = 11_e$

Второй субалфавит UC2:

- $CC = CG = AT = AA = 00_q$
- $CA = CT = AC = AG = 01_q$
- $GG = GC = TA = TT = 10_q$
- $GT = GA = TC = TG = 11_q$

Третий субалфавит UC3:

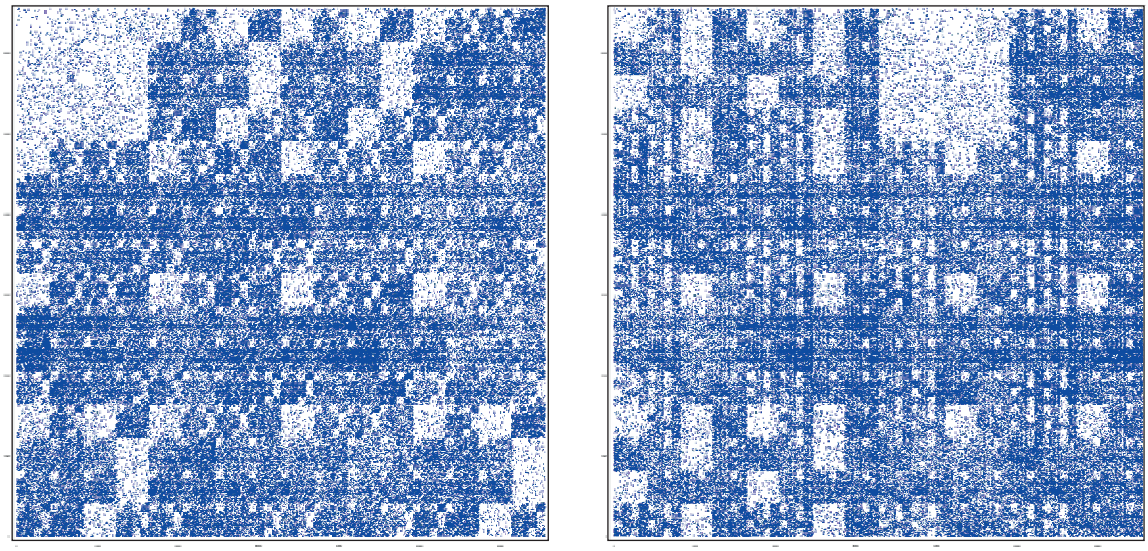
- $CC = GC = CA = GA = 00_s$
- $AC = TC = TA = AA = 01_s$
- $GG = CG = GT = CT = 10_s$
- $TG = AG = AT = TT = 11_s$

Четвертый субалфавит UC4:

- $CC = GG = AG = TC = 00_p$
- $AA = TT = CT = GA = 01_p$
- $AT = TA = CA = GT = 10_p$
- $CG = GC = AC = TG = 11_p$

Полученная нами визуализация генома китайского аллигатора обнаруживает его выраженный фрактальный характер (рис. 12). Этот подход существенно отличается от метода CGR, как алгоритмически, так и по генерируемым отображениям. Другие варианты параметризации

на основе различных декомпозиционных субалфавитов из матричной генетики требуют дополнительных исследований в рамках концепции геометрии. В частности, автором были предприняты попытки визуализации на основе матриц вырожденности аминокислот [Петухов, 2008], но эти визуализации пока не обнаружили четкой упорядоченности или фрактальных структур. Представления, построенные на основе субалфавитов для систем 2-блочных U-комплексных чисел из таблицы 1 также требуют исследований.



Геометрическая проекция UC2–UC4 Геометрическая проекция UC3–UC4

Рис. 12. Проекции ДНК китайского аллигатора (*Alligator sinensis*) при $N = 8$ и использовании субалфавитов систем 2-блочных *U*-комплексных чисел. Видны элементы фрактальной структуры с повторяющейся самоподобной мозаикой на различных масштабах. Ось абсцисс кодирует номер 8-плета по субалфавиту UC4 (слева и справа), ось ординат — по субалфавитам UC2 (слева) и UC3 (справа)

Предлагаемая автором концепция выдвигается в качестве нового направления в области разработки программного обеспечения для анализа биологической информации на молекулярно-генетическом и надмолекулярном уровне. В частности, изложенные методы позволяют «разглядеть» границу перехода между молекулярно-генетическим уровнем (на котором живая материя организуется в сложные высокомолекулярные органические соединения — белки, нуклеиновые кислоты и др.) и надмолекулярным (субклеточным) уровнем (на котором во внутриклеточных структурах хромосом и митохондрий уложены молекулы ДНК). Причем данная граница перехода обнаруживается в многомерном многомасштабном параметрическом пространстве с содержательным множеством биологически интерпретируемых проекций. Поэтому результаты применения изложенных алгоритмов позволяют говорить о новых методах исследования популяционно-видового уровня организации живых систем с применением геометрического подхода. При этом поиск систем субалфавитного кодирования для *n*-плетов, которые при визуализации дают фрактальные или упорядоченные паттерны, является одной из задач геометрии.

Как было отмечено, представленные алгоритмы реализованы в пространствах двоично-ортогональных функций Уолша и позволяют оценить виды соотношений между присутствующими и отсутствующими *N*-мерами в геномах различных организмов и вирусов (становится видно, что этим соотношениям свойственна фрактально-кластерная структура). Результатом физико-химической интерпретации первого правила Чаргаффа является модель двойной спирали Дж. Уотсона и Ф. Крика [Crick, 1979] (если комплементарные основания в целой молекуле ДНК идут всегда парами, то их количества одинаковы, что очевидно). Остальные правила Чар-

гаффа рассматриваются как феноменологические, но они позволяют рассматривать ДНК уже как геометрическую модель по системе бинарных субалфавитов нуклеотидов (а не двойной спирали) с визуализацией кодирования. В связи с этим результаты данного исследования позволяют выдвинуть разработанную биоматематическую систему методов описания нуклеиновых кислот в качестве дополнительной к структурной модели двойной спирали Дж. Уотсона и Ф. Крика.

Таким образом, предлагаемая в данной работе методология исследований способствует разработке инструментария для упрощения восприятия для последующего анализа и обобщения феномена генетического кодирования. Этот инструментарий был нами разработан и применен для построения геометрических отображений. При разработке изложенных в данной работе алгоритмов и соответствующего программного обеспечения принималось во внимание свойство зрительного анализатора человека, заключающееся в том, что двумерные объекты воспринимаются им наиболее эффективно: природа предусмотрела использование богатых математических свойств двумерия, которое положено в основу поверхности коры головного мозга (неокортекса), ответственного за анализаторные функции и высшую нервную деятельность. Неокортекс компактно свернут в черепной коробке, образуя извилины, количество которых (а значит, и площадь используемой двумерной поверхности) влияет на когнитивные способности. Результаты проведенных исследований показали, что именно двумерные геометрические представления обладают наиболее выраженными фрактальными и симметрическими свойствами.

## 7. Заключение

Представленная концепция выводит проблему компьютерного анализа генетической информации на принципиально новый уровень. Прежде всего изложенная система методов способствует оптимизации естественного интеллекта исследователя, усиливая его возможности, так как нуклеиновые кислоты имеют наглядные параметрические представления, позволяющие оценивать вариативность их физико-химических параметров визуально. Методы визуализации расширяют арсенал известных математических методов биоинформатики. CGR-метод оказывается частным случаем изложенных геометрических алгоритмов. Как следует из результатов исследования, вычислительные методы визуализации как целых молекул ДНК и РНК, так и их фрагментов обосновывают связь их физико-химических параметров с объектами дискретной геометрии. Это обстоятельство может помочь в исследовании внутренних симметрий, фрактальности и других характеристик нуклеиновых кислот, в том числе для изучения сложных взаимоотношений между различными видами живых организмов. Появление обоснованных методов сопоставления генетических моделей с теми или иными фенотипическими признаками способствует расширению методов исследований в биоинформатике и развитию численных методов, а также соответствующего научно-исследовательского программного обеспечения.

С точки зрения компьютерного моделирования представленная система методов перспективна в качестве математической основы для построения информативных и интерпретируемых моделей для внедрения в программное обеспечение биоинформационных СУБД. При этом возникает дополнительный функционал: новые инструменты аннотирования, разметки, геометрического поиска, сжатия и представления генетических последовательностей, позволяющие производить визуальный анализ молекулярно-биологической информации. Предлагаемый метод нацелен на решение следующих задач:

- визуализация и анализ нуклеотидного состава для исследования больших генетических данных (Big Data);
- исследование работающих в нуклеотидном составе феноменологических правил Чаргаффа, которые описывают комбинаторные варианты нуклеотидного состава при выполнении определенных ограничивающих соотношений;

- разработка инструментария хранения, поиска, классификации и кластеризации генетических данных по их биофизическим параметрам.

Подводя итог, следует отметить, что характерным отличительными чертами авторских алгоритмов являются: внутренняя биологическая интерпретируемость, которая закладывается на этапе параметризации, возможность задания параметров на основе элементов матричной генетики [Петухов, 2008] и возможности выбора модельного пространства (его размерности, вида и др.).

Использование численных методов в сравнительной геномике всегда было важно в биоинформатике. Изложенный подход позволяет увеличить эффективность исследований за счет доступности, наглядности и интерпретируемости, что открывает возможности подключения к исследованиям не только узких специалистов, но и всех интересующихся в рамках концепции краудсорсинга («привлечение к решению тех или иных проблем инновационной производственной деятельности широкого круга лиц для использования их творческих способностей, знаний и опыта по типу субподрядной работы на добровольных началах с применением инфокоммуникационных технологий» [Егерев, Захарова, 2013]). Это становится все более актуальным в связи с ростом количества секвенированных (оцифрованных) генетических данных.

## Список литературы (References)

*Егерев С. В., Захарова С. А.* Краудсорсинг в науке // Наука. Инновации. Образование: альманах / Российский научно-исследовательский ин-т экономики, политики и права в научно-технической сфере (РИЭПП). — Языки славянской культуры. — 2013. — № 14. — С. 175–186. — ISSN 1996-9953.

*Egerev S. V., Zakharova S. A.* Kraudsorsing v nauke [Crowdsourcing in science] // Al'manakh “Nauka. Innovatsii. Obrazovanie” / Rossiiskii nauchno-issledovatel'skii in-t ekonomiki, politiki i prava v nauchno-tekhnicheskoi sfere (RIEPP) [Almanac “Science. Innovation Education”/ Russian Research Institute of Economics, Politics and Law in the Scientific and Technical Field (RIEPP)]. — Yazyki slavyanskoj kul’tury [Languages of Slavic culture]. — 2013. — No. 14. — P. 175–186 (in Russian).

*Петухов С. В.* Матричная генетика, алгебры генетического кода, помехоустойчивость. — М.: РХД, 2008. — 316 с.

*Petukhov S. V.* Matrichnaya genetika, algebry geneticheskogo koda, pomekhoustoichivost' [Matrix genetics, algebras, genetic codes, noise immunity]. — Moscow: RCD, 2008. — 316 p. (in Russian).

*Петухов С. В.* Гиперкомплексные числа и алгебраическая система генетических алфавитов // Гиперкомплексные числа в геометрии и физике. — 2011. — Т. 8, № 2 (16). — С. 118–138.

*Petukhov S. V.* Giperkompleksnye chisla i algebraicheskaya sistema geneticheskikh alfavitov [Hypercomplex numbers and the algebraic system of genetic alphabets] // Giperkompleksnye chisla v geometrii i fizike [Hypercomplex numbers in geometry and physics]. — 2011. — Vol. 8, No. 2 (16). — P. 118–138 (in Russian).

*Петухов С. В.* Гиперкомплексные числа, генетическое кодирование и алгебраическая биология // Метафизика. — 2012. — № 3 (5). — С. 64–88.

*Petukhov S. V.* Giperkompleksnye chisla, geneticheskoe kodirovanie i algebraicheskaya biologiya [Hypercomplex numbers, genetic coding and algebraic biology] // Metafizika [Metaphysics]. — 2012. — No. 3 (5). — P. 64–88 (in Russian).

*Румер Ю. Б.* Систематизация кодонов в генетическом коде // Доклады АНСССР. — 1968. — Т. 183, № 1. — С. 225–226.

*Rumer Yu. B.* Sistematizatsiya kodonov v geneticheskom kode [Systematization of codons in the genetic code] // Doklady ANSSSR [Reports of the ANSSSR]. — 1968. — Vol. 183, No. 1. — P. 225–226 (in Russian).

*Сагалович Ю. Л.* Введение в алгебраические коды: учебное пособие. — М.: МФТИ, 2007. — 262 с.

*Sagalovich Yu. L.* Vvedenie v algebraicheskie kody: Uchebnoe posobie [Introduction to Algebraic Codes: a Textbook]. — Moscow: MIPT, 2007. — 262 p. (in Russian).

*Седов А. Е.* Метафоры в генетике // Вестник РАН. — 2000. — Т. 70, № 6.

*Sedov A. E.* Metafory v genetike [Metaphors in genetics] // Vestnik RAN [Bulletin of the Russian Academy of Sciences]. — 2000. — Vol. 70, No. 6 (in Russian).

*Хармут Х.* Применение методов теории информации в физике. — М.: Мир, 2016. — 344 с.
*Harmut X.* Primenenie metodov teorii informatsii v fizike [Application of information theory methods in physics]. — Moscow: Mir, 2016. — 344 p. (in Russian).

*Axelrod R. M.* The Structure of Decision: Cognitive Maps of Political Elites. — Princeton University Press, 1976.

*Balonin N. A., Balonin Y. N., Djokovic D. Z., Karbovskiy D. A., Sergeev M. B.* Construction of symmetric Hadamard matrices. — https://arxiv.org/abs/1708.05098

*Barnsley M.* Fractals Everywhere. — New York: Academic Press, 1988. — 534 p.

*Duarte-Sanchez J. E., Velasco-Medina J., Moreno P. A.* Hardware Accelerator for the Multifractal Analysis of DNA Sequences // IEEE/ACM Trans Comput Biol Bioinform. — 2018. — Vol. 15, No. 5. — P. 1611–1624.

*Eidenzon D., Pilipczuk O.* Multidimensional Data Visualization // Encyclopedia of Information Science and Technology. — Third Edition. — 2015. — DOI: 10.4018/978-1-4666-5888-2.ch153

*Chargaff E., Lipshitz R., Green C.* Composition of the deoxypentose nucleic acids of four genera of sea-urchin // J. Biol. Chem. — 1952. — Vol. 195, No. 1. — P. 155–160. — PMID 1493836.

*Crick F. H., Wang J. C., Bauer W. R.* Is DNA really a double helix? // J. Mol. Biol. — April 1979. — Vol. 129, No. 3. — P. 449–57. — doi:10.1016/0022-2836(79)90506-0. — PMID 458852.

*Feldman D. P.* 17.4 The chaos game // Chaos and Fractals: An Elementary Introduction. — Oxford University Press, 2012. — P. 178–180. — ISBN 9780199566440.

*Stepanyan I. V., Petoukhov S. V.* The Matrix Method of Representation, Analysis and Classification of Long Genetic Sequences // Information. — 2017. — Vol. 8. — P. 12.

*Hewelt B., Li H., Jolly M. K., Kulkarni P., Mambetsariev I., Salgia R.* The DNA walk and its demonstration of deterministic chaos-relevance to genomic alterations in lung cancer // Bioinformatics. — 2019. — Vol. 35 (16). — P. 2738–2748.

*Hutchinson J.* Fractals and self similarity // Indiana University Mathematics Journal. — 1981. — Vol. 30, No. 5. — P. 713–747.

*Tikhomirov G., Petersen P., Qian L.* Fractal assembly of micrometre-scale DNA origami arrays with arbitrary patterns // Nature. — 2017. — Vol. 552 (7683). — P. 67–71.

*Townsend J. P., Su Z., Tekle Y.* Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny // Genetics. — 2012. — Vol. 61 (5). — P. 835–849. — doi:10.1093/sysbio/sys036. PMID 22389443

*Pei S., Dong W., Chen X., He R. L., Yau S. S.* Fast and accurate genome comparison using genome images: The Extended Natural Vector Method // Mol. Phylogenet. Evol. — 2019. — Vol. 141. — P. 106633.

*Petoukhov S. V., He M.* Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications. — Hershey, USA: IGI Global, 2010. — 271 p.

*Petoukhov S. V.* Genetic coding and united-hypercomplex systems in the models of algebraic biology // Biosystems. — 2017. — Vol. 158. — P. 31–46.

*Yin C.* Encoding and Decoding DNA Sequences by Integer Chaos Game Representation // J. Comput. Biol. — 2019. — Vol. 26 (2). — P. 143–151.

*Zenkin A. A.* Cognitive computer graphics — application to decision support systems // Proc of II Int. Conf MORINTECH–97. — St. Petersburg, Russia, 1997. — Vol. 8. — P. 197–203.