Научная визуализация, 2020, том 12, номер 3, страницы 61 - 78, DOI: 10.26583/sv.12.3.06

Масштабно-параметрическая визуализация нуклеиновых кислот

И.В. Степанян[1]

Институт машиноведения им. А.А. Благонравова Российской академии наук

[1] ORCID: 0000-0003-3176-5279, neurocomp.pro@gmail.com

Аннотация

Работа описывает новые результаты в области алгебраической биологии, где используются матричные методы (Петухов, 2019, 2012, 2008; Петухов, Не, 2010) с переходом от матричной алгебры к конечной геометрии и компьютерной визуализации. Предложенный метод масштабно-параметрической визуализации нуклеиновых кислот позволяет отобразить биохимический состав нуклеотидных последовательностей в пространствах двоично-ортогональных функций Уолша, кодирующих физико-химические параметры нуклеотидов. Приведены примеры визуализации нуклеотидного состава геномов различных видов живых организмов. Результаты визуализации сопоставлены с системой феноменологических правил Чаргаффа, описывающих количественные соотношения между различными типами азотистых оснований ДНК. В результате проведённых исследований установлено, что разработанный метод может служить для упрощения восприятия длинных нуклеотидных последовательностей путем их визуализации в пространствах различной размерности, а также служить дополнительным критерием классификации и выявления межвидовых взаимосвязей. Также установлено, что предложенный метод позволяет обосновать связь параметров молекул ДНК и РНК с фрактальными геометрическими мозаиками, обнаруживает упорядоченность и симметрии полинуклеотидов и помехоустойчивость их визуальных представлений.

Ключевые слова: методы визуализации, функции Уолша, правила Чаргаффа, многомерный анализ, нуклеотидный состав, конечная геометрия, фракталы, ДНК, хромосомы, симметрии.

1. Введение. Общие сведения о нуклеотидах и нуклеотидном составе ДНК и РНК

Нуклеиновые кислоты ДНК и РНК – последовательности комплементарных пар нуклеотидов, выполняющие функции хранения и передачи наследственной генетической информации в живых организмах [1,17]. Эти последовательности анализируются, как правило, статистическими методами. Они имеют одномерный линейный характер и отображаются в виде строк, состоящих из четырех букв алфавита, кодирующего нуклеотиды: аденин (A), гуанин (G), цитозин (C) и тимин (T) (урацил (U)).

Визуальный анализ длинных сток, состоящих из букв, кодирующих нуклеотиды реальных генетических последовательностей является трудоёмкой задачей. Для её упрощения разработано множество алгоритмов и программных продуктов, позволяющих визуализировать и анализировать ДНК с применением различных гистограмм, таблиц и графиков, см. например [23-26]. Эти методы основаны на машинном статистическом анализе и широко применяются в научных исследованиях. В данной работе нами была поставлена задача по разработке нового метода, который упрощает визуальный

анализ длинных нуклеотидных последовательностей (вопрос интерпретации нуклеотидного состава лежит за пределами данного исследования).

2. Материалы и методы. Кодирование физико-химических параметров нуклеотидов в симметричной матрице Адамара

2.1 Генетическое кодирование как система функций Уолша

В [2] показано, что каждое азотистое основание генетического кода имеет три варианта своего двоичного представления. Эти варианты представлений, названные С.В. Петуховым бинарными суб-алфавитами, различаются в соответствии с типами бинарно-оппозиционных свойств в наборе азотистых оснований:

- G = C «3 водородные связи» / A = T «2 водородные связи»;
- C = T «пиримидины» / A = G «пурины»;
- A = C «амино» / G = T «кето» [20];
- A=T=G=C (наличие фосфатного остатка).

С учетом дополнительного четвертого признака, который не является оппозиционным, система генетических суб-алфавитов может быть представлена в виде матрицы Адамара, показанной на рис. 1.

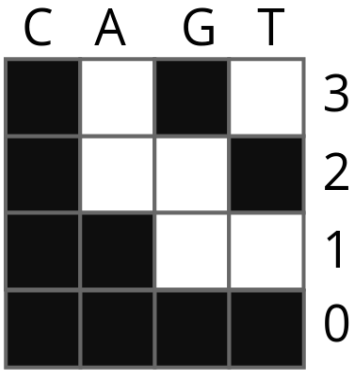


Рис. 1. Вариант матрицы Адамара, отображающей кодирование нуклеотидных суб-алфавитов. Затемненные клетки +1, белые клетки –1 (или наоборот в зависимости от способа кодирования). Номера суб-алфавитов обозначены как 0, 1, 2 и 3.

Данная матрица является симметричной, поскольку нуклеотиды могут быть заменены соответствующими суб-алфавитами без изменения структуры матрицы (строки и столбцы можно менять местами) [3]. Каждая строка и каждый столбец матрицы Адамара является функцией Уолша [4]. Функции Уолша – это полный набор ортогональных функций, которые могут быть использованы для представления любой дискретной функции по аналогии с использованием в Фурье-анализе тригонометрических функций [7]. Они используются в цифровой технике, при кодировании помехоустойчивой связи, в квантовой информатике и квантовой механике.

2.2 Генетическое кодирование как система правил Чаргаффа

Э. Чаргафф выявил систему биохимических закономерностей внутри последовательностей нуклеиновых кислот, которая описывает количественные соотношения между различными типами нуклеотидов [1]. Эта система закономерностей представляет собой набор алгебраических соотношений:

1. Количество аденина равно количеству тимина, гуанина – цитозину:
   A = T, G = C или A / T = 1, G / C = 1 (пары Уотсона-Крика [17]).

2. Количество пуринов равно количеству пиримидинов:
   A+G ≈ T+C или (A+G) / (C+T) ≈ 1.
3. Число оснований с аминогруппами в положении 6 равно числу оснований с кетогруппами в том же положении:
   A+C ≈ T+G или (A+C) / (G+T) ≈ 1.
4. Соотношение (A+T) / (G+C) является коэффициентом специфичности и может быть различным с преобладанием пар AT или GC в зависимости от того или иного вида организма, реализуя многообразие живых форм.

Как видно из изложенного, нуклеотидная последовательность живого организма – сбалансированная система, представляющая собой двойную спираль (ДНК) и обладающая внутренними симметриями и определёнными математическими закономерностями. Дополнительная информация о симметриях и матрицах Адамара в генетическом кодировании, а также о генетических алгебрах подробно изложена в трудах биоматематика С.В. Петухова [2,10,15].

В связи с существованием связи между алгеброй и геометрией (а значит существованием связи между генетическими алгебрами и генетическими геометриями), автором была поставлена и решена задача разработки метода визуализации нуклеиновых кислот. Исследование строилось на гипотезе, что визуализация должна отображать симметрии нуклеотидного состава. Авторский метод позволяет исследовать феномен генетического кодирования с геометрической стороны.

2.3 Метод масштабно-параметрической визуализации нуклеиновых кислот

Приведенный метод представляет собой алгоритм компьютерной обработки биологической информации для масштабно-параметрической визуализации нуклеиновых кислот в координатных пространствах различной размерности. Основные идеи данного метода были впервые предложены автором в [5]. Далее приведены шаги разработанного алгоритма.

1) Масштабирование. Последовательность символов {A,G,T,C}, кодирующих азотистые основания в нуклеиновой кислоте, разделяется на фрагменты равной длины N, где N – свободный параметр алгоритма. Полученные фрагменты равной длины будем называть N-мерами или N-плетами [5].

2) Параметризация. С учетом системы генетических суб-алфавитов 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 повороты и другие преобразования итоговой визуализации (поэтому для возможности адекватного сопоставления полученных результатов необходимо проводить исследования с привязкой к «единому стандарту кодирования»).

3) Геометризация. Бинарная запись фрагментов является их представлением в виде трех последовательностей десятичных или иных однозначно-идентифицирующих значений. Преобразование двоичных N-меров в десятичные числа позволяет их отобразить в любой координатной системе. Числовые значения задают координаты точек в пространстве параметров (далее – в пространстве визуализации или параметрическом пространстве).

Примечание 1. 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 параметрических пространствах.

Примечание 2. Шаги 1 и 2 могут быть переставлены (сначала параметризация, затем масштабирование), что влияет на вычислительную нагрузку при расчёте длинных генетических последовательностей на ЭВМ.

Алгоритм визуализации был реализован автором в виде библиотеки программ на языках программирования Python, Lua, Moonscript и C++ без интерактивного редактора (GUI) и аппаратного ускорения графики. Были использованы специализированные модули для ускорения расчётов. Среднее время, требуемое на обработку генетической информации составляет от нескольких секунд до нескольких часов в зависимости от масштаба N и длины анализируемой последовательности. Иногда приходилось останавливать счет из-за превышения допустимого интервала времени. Часть расчетов была выполнена на суперкомпьютере «МВС-10П» (МСЦ РАН).

Предлагается эвристическая формула для вычисления масштаба N при визуализации нуклеотидных последовательностей длины L:

$$N = \left\lfloor \log_2\left(\sqrt{L}\right) \right\rfloor,$$

где квадратные скобки – операция взятия целой части числа. Ширина и высота квадратного изображения в точках:

$$K = 2^N$$

В качестве двумерных пространств проецирования предлагается выбирать все три возможные варианта комбинаций пар базисных функций Уолша. При этом, наиболее информативный вариант комбинации этих функций может зависеть от конкретного вида организма. На данный момент представляется, что существуют формальные правила по выбору базисных функций, но этот вопрос требуется дополнительно изучить, проанализировав предложенным способом строение большого количества ДНК разных видов организмов.

Метод относится к развитию статистических методов анализа нуклеотидных последовательностей и основан на параметризации, масштабировании и геометризации физико-химических параметров молекулы. В результате применения метода задаётся параметрическое пространство, которое является конечным, дискретным и трехмерным по количеству бинарно-оппозиционных признаков. 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 суб-алфавитов. Метод позволяет визуализировать нуклеотидный состав в различных проекциях, с различным масштабированием и по различным суб-алфавитам и может быть применен для анализа молекул РНК и ДНК.

Предлагаемая методика проведения исследований с применением разработанного метода: построение на основе предложенного метода примеров визуализации длинных нуклеотидных последовательностей из ДНК различных организмов:

- в трёхмерном пространстве физико-химических параметров, которое задаётся тремя строками 1,2 и 3 матрицы Адамара на рис. 1;
- в трёх двумерных пространствах физико-химических параметров, которые задаются тремя возможными вариантами сочетаний строк 1-2, 1-3 и 2-3 матрицы Адамара на рис. 1;
- в трёх одномерных пространствах физико-химических параметров, которые задаются тремя строками 1,2 и 3 матрицы Адамара на рис. 1, рассматриваемые по отдельности и по всей длине молекулы, что позволяет учитывать местоположение N-меров в генетической последовательности;
- нулевая (нижняя) строка матрицы Адамара на рис. 1 не является информативной, так как не кодирует бинарно-оппозиционных признаков, поэтому не рассматривается;
- дополнительно в соответствии с теорией секвентного анализа Хармута [6] возможны визуализации 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 парамет-

ров в *N*-плете, соответствующие пространства визуализации будем называть интегральными.

В ходе исследований были построены визуальные паттерны около сотни геномов простейших, растений, грибов, животных и вирусов. В данной работе для визуализации использовались геномы из биоинформационной базы данных NCBI [14], а также материалы, любезно предоставленные лабораторией проф. Н.С. Зенкина Центра биологии бактериальной клетки университета Ньюкасла (Великобритания).

3. Результаты и обсуждение. Примеры визуализации нуклеиновых кислот в параметрических пространствах различной размерности

3.1 Визуализация нуклеиновых кислот в трехмерном пространстве физико-химических параметров нуклеотидов

Ортогональный базис {X, Y, Z}, выбранный в качестве трехмерной декартовой системы координат, дает визуализацию, пример которой показан на рис. 2. Каждой точке соответствует обобщенная характеристика учитываемых бинарно-оппозиционных признаков соответствующего фрагмента последовательности, что позволяет отобразить нуклеотидный состав молекулы.

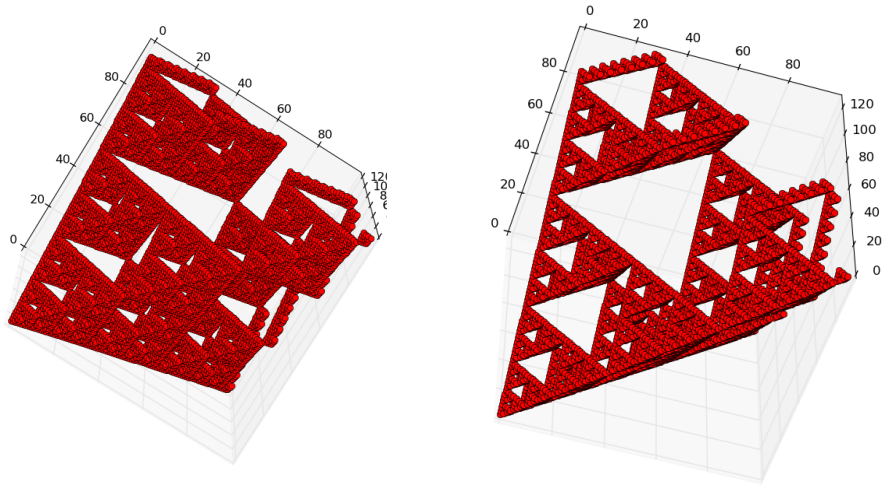
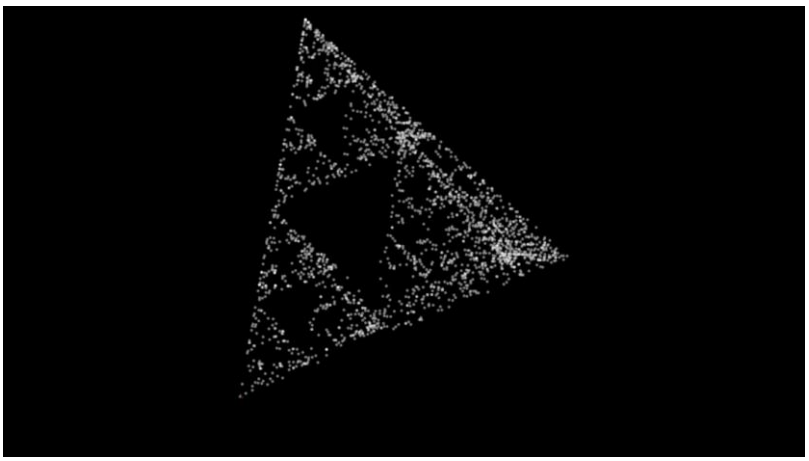


Рис.2. Иллюстрации трехмерного представления нуклеотидного состава на примере хромосомы живого организма в различных проекциях, построенная по авторскому алгоритму. Осям X, Y и Z соответствуют упорядоченные по возрастанию десятичные представления двоичного кодирования *N*-плетов на основе всех трех бинарно-оппозиционных суб-алфавитов. Каждой точке фигуры соответствует *N*-мер, координата которого задана его протонно-числовыми характеристиками. Анализ данных трехмерных изображений затруднен из-за геометрии самого объекта. Для устранения этого затруднения необходимо строить двумерные проекции.

Результирующая геометрическая фигура, напоминающая «симплекс Серпинского», типична для трехмерной визуализации любых длинных нуклеотидных последовательностей. Форма фигуры обусловлена свойствами бинарных суб-алфавитов и свойствами матрицы Адамара на рис. 1. Координаты каждой точки в трехмерном пространстве визуализации задаются любой парой её координат, поскольку третья координата вычисляется путем сложения по модулю два двух оставшихся координат. Данная алгебраическая особенность связана с избыточностью бинарных суб-алфавитов, использующихся для хранения и передачи генетической информации по цепям поколений. Анимированный вариант рис. 2 представлен в анимации:



Пример интегрального трехмерного представления нуклеотидного состава хромосомы живого организма приведен на рис. 3. Это объект конечной геометрии, каждая точка которого соответствует множеству *N*-меров нуклеиновой кислоты, объединенных по числу единиц в двоичном кодировании.

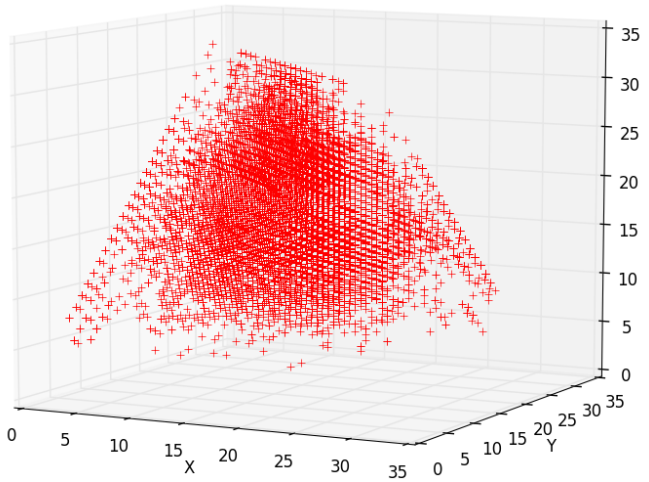
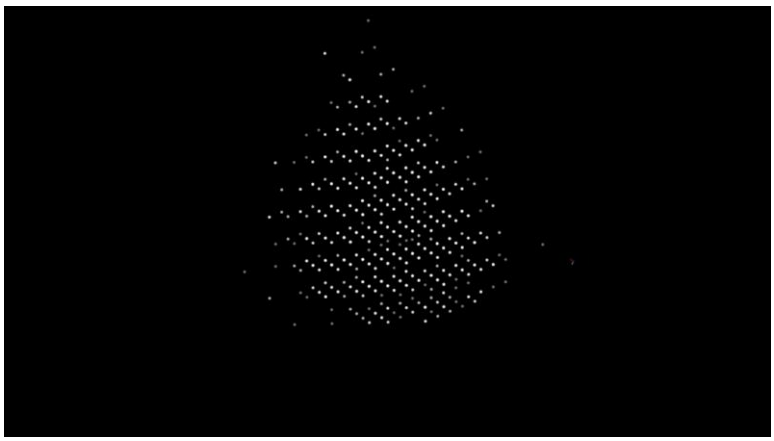


Рис.3.Интегрально-трехмерное представление нуклеотидного состава хромосомы. Осям X, Y и Z соответствуют количества единиц в десятичных представлениях двоичного кодирования каждого *N*-плета, используя три бинарно-оппозиционных субалфавита.

Анимированный вариант рис. 3 представлен в анимации:



3.2 Визуализация нуклеиновых кислот в трех двумерных пространствах физико-химических параметров нуклеотидов

Свойства параметрического пространства таковы, что трехмерные представления не удобны для восприятия и анализа особенностей длинных нуклеиновых кислот. Однако, двумерные проекции этого трехмерного представления подходят для отображения специфики их строения. В базисах {X, Y}, (X, Z} и {Y, Z}, выбранных в качестве декартовых систем координат, трехмерная визуализация дает три различные двумерные проекции на основе соответствующих суб-алфавитов физико-химических параметров нуклеотидов.

На основе разработанного метода визуализации и компьютерной программы обнаружено, что хромосомы различных видов организмов имеют индивидуальные особенности строения. Визуализация геномов различных организмов может иметь двумерный паттерн, который визуально подобен для всех хромосом и их произвольных фрагментов, а также для всего рассматриваемого организма. На рис. 4-9 приведены примеры двумерной визуализации различных нуклеотидных последовательностей. Рядом с рисунками в порядке A, G, T, C приведены пары функций Уолша, которые использовались для кодирования их физико-химических параметров (строки матрицы Адамара из рис. 1).

Исходя из отмеченного свойства генетического кодирования (согласно которому тройка бинарно-оппозиционных суб-алфавитов связана между собой операцией сложения по модулю 2) для определения произвольной нуклеиновой кислоты достаточно любой пары бинарных представлений. Поэтому для двумерной визуализации нуклеотидного состава достаточно любой пары осей координат. Как оказалось, вопрос определения наиболее информативной пары координатных осей (и соответственно учитываемых параметров) зависит от вида живого организма.

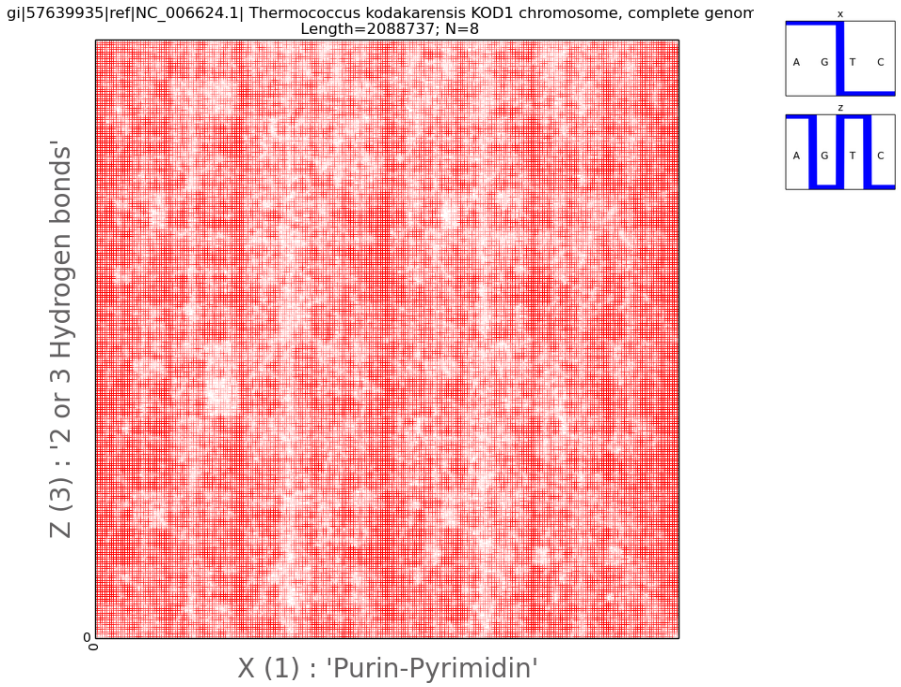


Рис.4. Иллюстрация двумерного представления нуклеотидного состава хромосомы термофильной археи. Пара Уолш-функций, используемых для параметризации отображается в верхнем правом углу. Осям абсцисс и ординат соответствуют десятичные представления двоичного кодирования каждого 8-плета.

В результате проведенного анализа было обнаружено, что из трех вариантов двумерной визуализации часто наиболее информативными и симметричными являются

мозаики на основе информации о внешнем строении молекулы, т.е. построенные на элементах структур, кодирующих признаки амино/кето и пурин/пиримидин. Такие мозаики имеют детализированный узор, в котором как правило прослеживаются прямоугольные формы (рис. 4, 7-9). Однако, в некоторых случаях наиболее выраженными и симметричными оказываются мозаики на основе типов водородных связей, отображающих внутреннюю структуру двойной спирали ДНК. Такие мозаики, как правило, характеризуется выраженными диагональными элементами паттерна и встречается например в ДНК митохондрии растения *arabidopsis thaliana* (рис. 5).

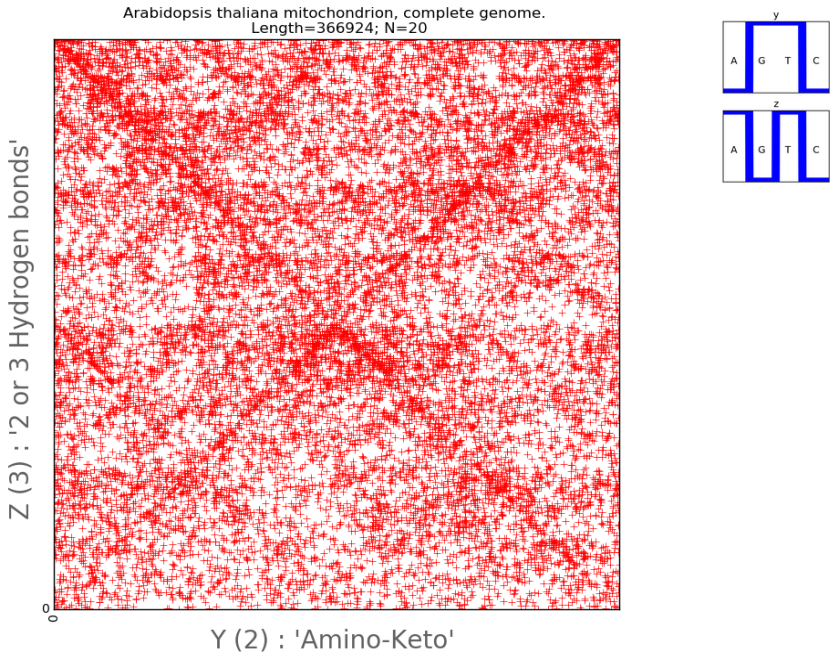


Рис.5. Иллюстрация двумерного представления нуклеотидного состава генома митохондрии растения Резуховидка (резушка) Таля (лат. *Arabidopsis thaliana*) семейства капустные (*Brassicaceae*). Пара Уолш-функций, используемых для параметризации, отображается в верхнем правом углу. Осям абсцисс и ординат соответствуют десятичные представления двоичного кодирования каждого 8-плета.

На рис. 6 и 7, где показана мозаика, отражающая внутреннее строение хромосом двух организмов, хорошо прослеживаются диагональные элементы. В геноме бактерии на рис. 6 видны фрактальные повторения диагоналей по всему паттерну. Диагональные элементы отличаются по цвету в зависимости от направления и места на фрактальном узоре. На рис. 7 визуализация нуклеотидного состава второй хромосомы одноклеточного микроскопического грибка «пекарские дрожжи» демонстрирует иное поведение диагональных элементов: диагонали хорошо прослеживаются только в одном направлении, фрактальные повторения диагоналей выражены также только в одном направлении, при этом они отображают отсутствующие *N*-меры. Противоположные диагонали, ответственные за присутствующие *N*-меры прослеживаются менее чётко.

Отметим, что диагональные и другие элементы узора могут быть направленными в различные стороны у разных организмов при сохранении общего строения узора. Данная особенность может быть смоделирована прочтением комплементарной нити ДНК.

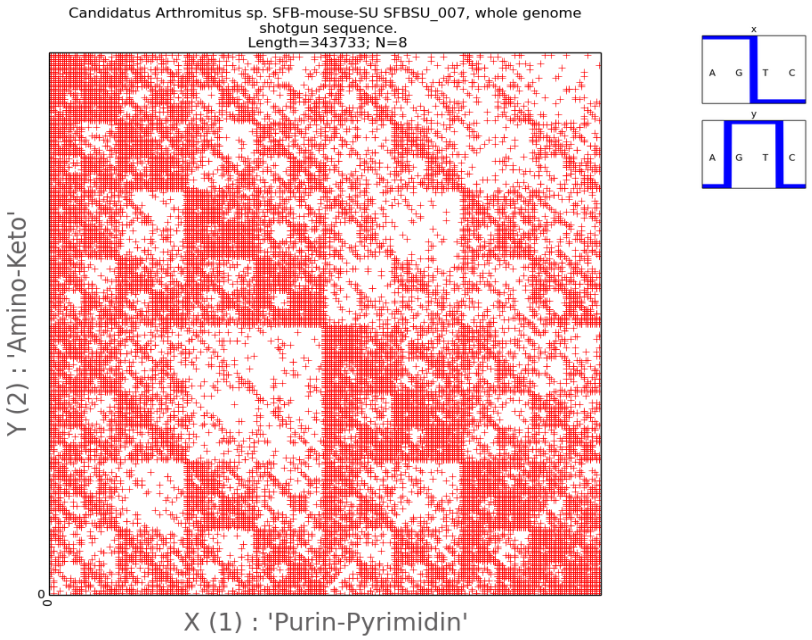


Рис.6. Иллюстрация двумерного представления нуклеотидного состава генома бактерии. Пара Уолш-функций, используемых для параметризации, отображается в верхнем правом углу. Осям абсцисс и ординат соответствуют десятичные представления двоичного кодирования каждого 8-плета.

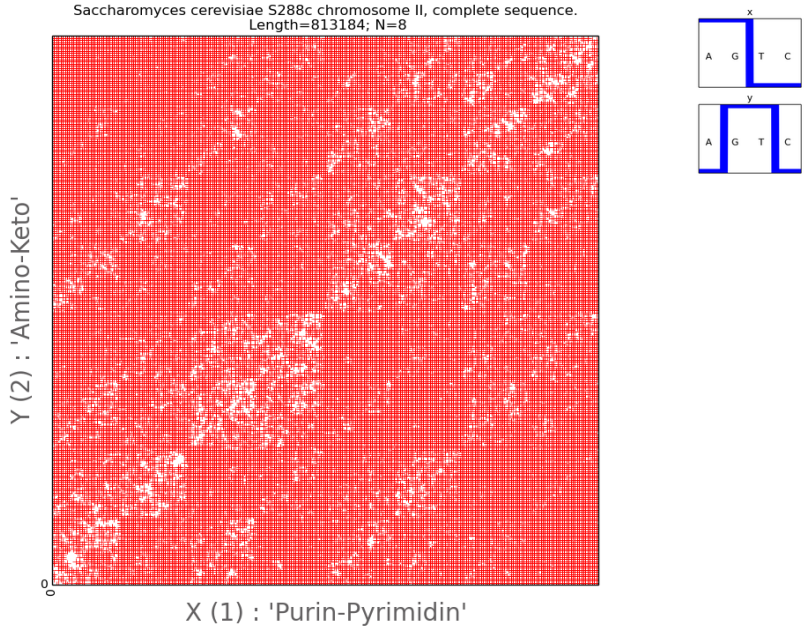


Рис.7. Иллюстрация двумерного представления нуклеотидного состава второй хромосомы одноклеточного микроскопического грибка (пекарские дрожжи). Пара Уолш-функций, используемых для параметризации, отображается в верхнем правом углу. Осям абсцисс и ординат соответствуют десятичные представления двоичного кодирования каждого 8-плета.

На рис. 8 и 9 приведены визуальные двумерные представления бактерии *Ralstonia eutropha (H16 megaplasmid pHG1)* и полного генома протобактерии *Burkholderia multivorans* соответственно. Их визуальные паттерны характеризуются выраженной фрактальностью, причём паттерн протобактерии обладает яркой формой – хорошо виден баланс присутствующих и отсутствующих 63-меров в её ДНК (рис. 9).

Общенаучные методы изучения нуклеиновых кислот, как правило, концентрируют свое внимание на те фрагменты, которые в них присутствуют. Предложенный метод позволяет представить в наглядной форме феноменологию и особенности дефицита и присутствия различных типов *N*-меров. Отсутствующие и присутствующие *N*-меры генома протобактерии на рис. 9 составляют красивый фрактал. Таким образом, геометрический подход позволяет отобразить баланс присутствующих и отсутствующих 63-меров, формирующих структурированные фрактальные кластеры на рис. 8 и 9.

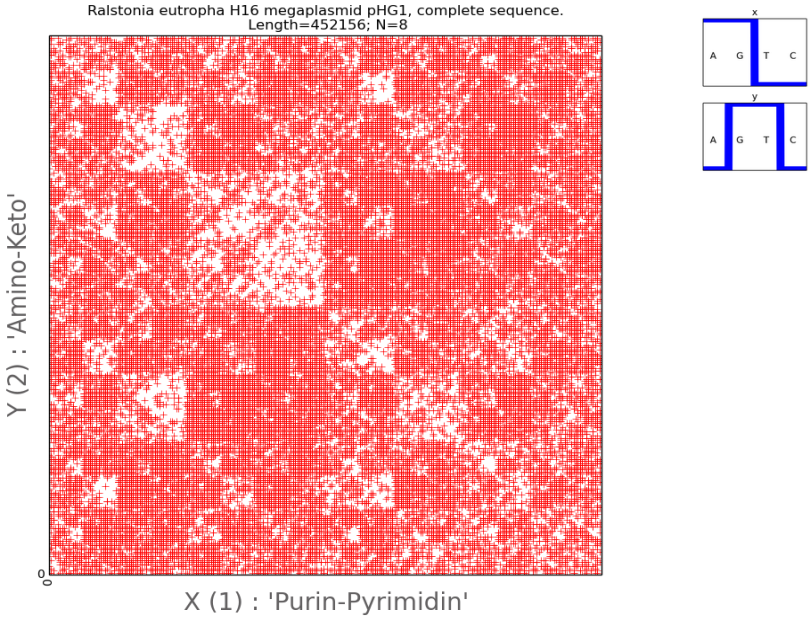


Рис.8. Иллюстрация двумерного представления нуклеотидного состава генома бактерии. Осям абсцисс и ординат соответствуют десятичные представления двоичного кодирования каждого 8-плета. Один из характерных паттернов, имеющих фрактальный характер.

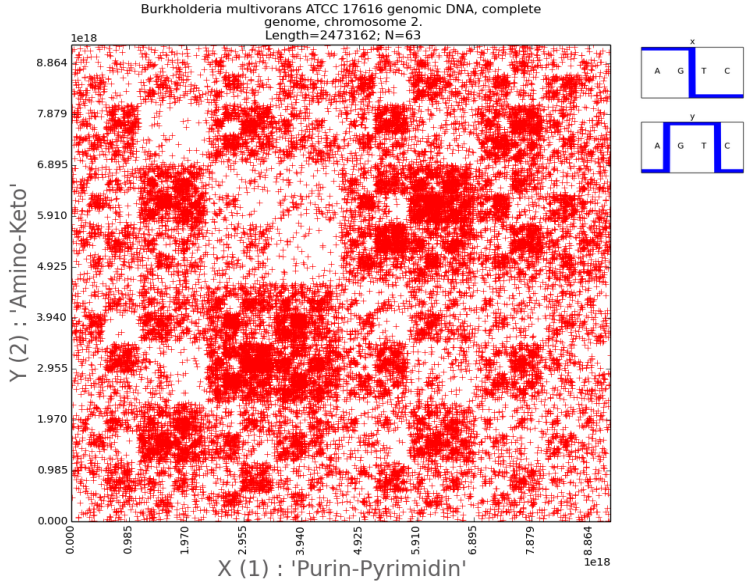


Рис.9. Иллюстрация двумерного представления нуклеотидного состава второй хромосомы протобактерии. Видно, что у данного организма присутствующие и отсутствующие 63-плеты образуют симметричную фрактальную мозаику, структура которой стабильна относительно реверсирования цветов. Осям абсцисс и ординат соответствуют десятичные представления двоичного кодирования каждого 63-плета.

Проведенные исследования и анализ визуализаций нуклеотидных последовательностей различных видов живых организмов подтверждает, что нуклеотидный состав может быть идентичным у организмов, которые не являются родственными в филогенетическом древе и различным у родственных организмов [12]. Известен специальный класс симметрий, реализованных в длинных ДНК-последовательностях разных организмов. В работе С.В. Петухова [22] приведены фрактальные генетические сети и описаны тетрагрупповые симметрии. Таким образом, известные научные данные о фрактальности ДНК наглядным образом отображаются в визуальном представлении на основе предложенного метода.

На рис. 10 приведен пример интегрально-двумерного представления нуклеотидного состава хромосомы человека на одной из плоскостей визуализации. Примеры генетических мозаик, построенных в непозиционной системе счисления приведены в [9].

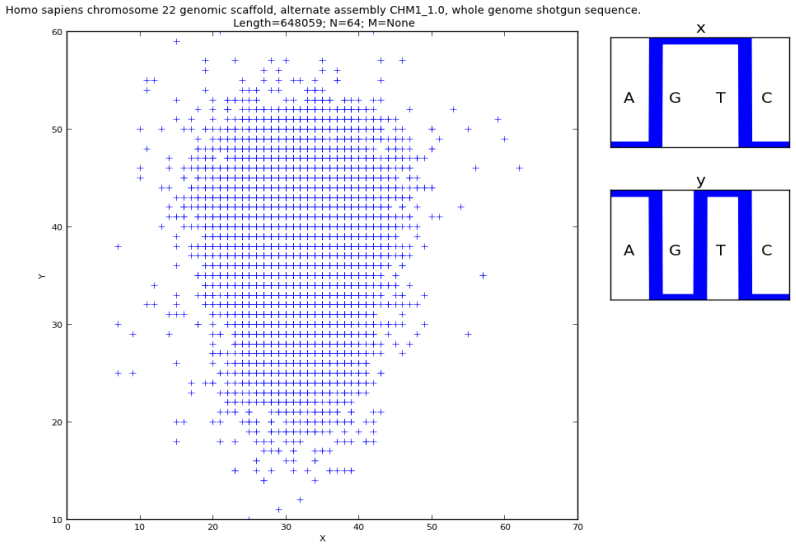
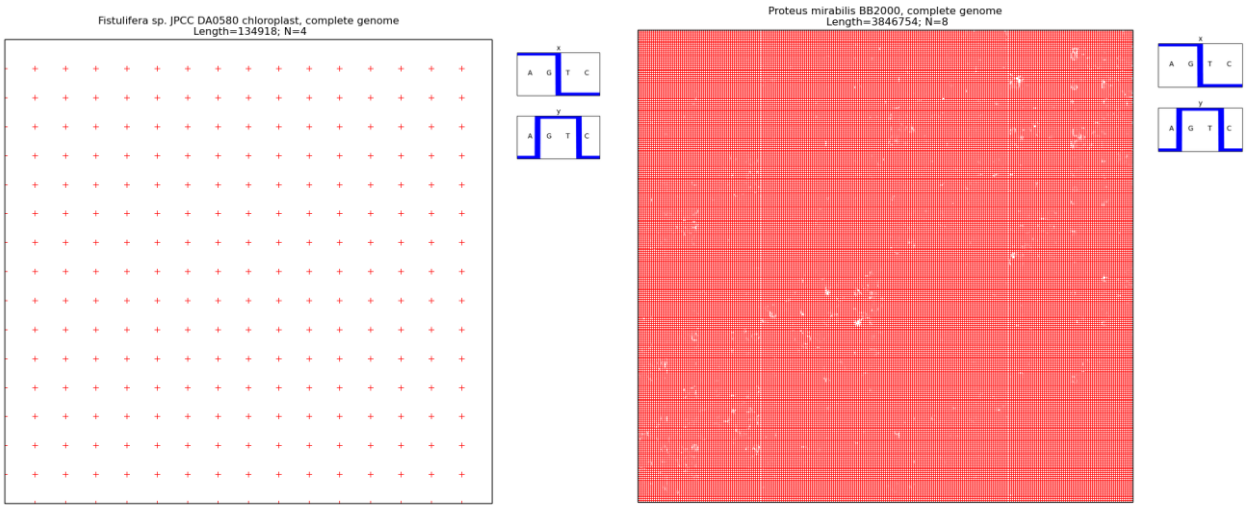


Рис.10. Иллюстрация интегрально-двумерного представления нуклеотидного состава хромосомы человека на одной из плоскостей визуализации. Пара Уолш-функций, используемых для параметризации отображена справа. Осям абсцисс и ординат соответствует количество единиц каждого 64-плета с использованием пары двоично-оппозиционных суб-алфавитов.

Предварительные результаты применения метода 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 (прореживания), 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, стабильность мозаичных узоров наблюдалась при удалении каждого второго нуклеотида, каждого третьего нуклеотида и т.д. При этом, визуализация нуклеиновых кислот в двумерных пространствах 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 определенного диапазона – этот эффект прослеживается в анимациях:



Для дальнейших исследований с помощью разработанной компьютерной программы были созданы случайные последовательности азотистых оснований длиной в 100000 нуклеотидов с разделением на *N*-плеты по 8, 16 и 28. Генерированные случайным образом последовательности при визуализации дали паттерн, все точки которого разбросаны хаотично (рис. 11, верхний ряд). Их визуальные представления носят нерегулярный, хаотический характер при полном отсутствии каких-либо мозаик по всем суб-алфавитам, что значительно отличает их от реальных длинных нуклеотидных последовательностей.

Также на компьютере нами были созданы псевдослучайные последовательности нуклеотидов с соблюдением второго правила Чаргаффа, действительного для каждой из двух нитей ДНК [1,11]. На рис. 11 приведено сравнение последовательностей, которые были случайным образом созданы без соблюдения (нижний ряд) и с соблюдением (верхний ряд) второго правила Чаргаффа. Для этих последовательностей был визуализирован специальный вид закономерностей при различных значениях *N*, равных 6,7 и 20. Из рис. 11 видно, что случайный паттерн, построенный по второму правилу Чаргаффа, структурирован за счёт наличия в нём пустых ровных областей, которые равномерно распределены и особенно чётко прослеживаются при *N*=6 на рис. 11 в нижнем ряду слева. В тоже время, как отмечено выше, случайный паттерн, созданный без соблюдения соотношения Чаргаффа имеет хаотический характер при визуализации (верхний ряд). Из этого можно сделать вывод о геометрической связи паттернов визуализации по авторскому алгоритму с алгебраическими правилами Чаргаффа.

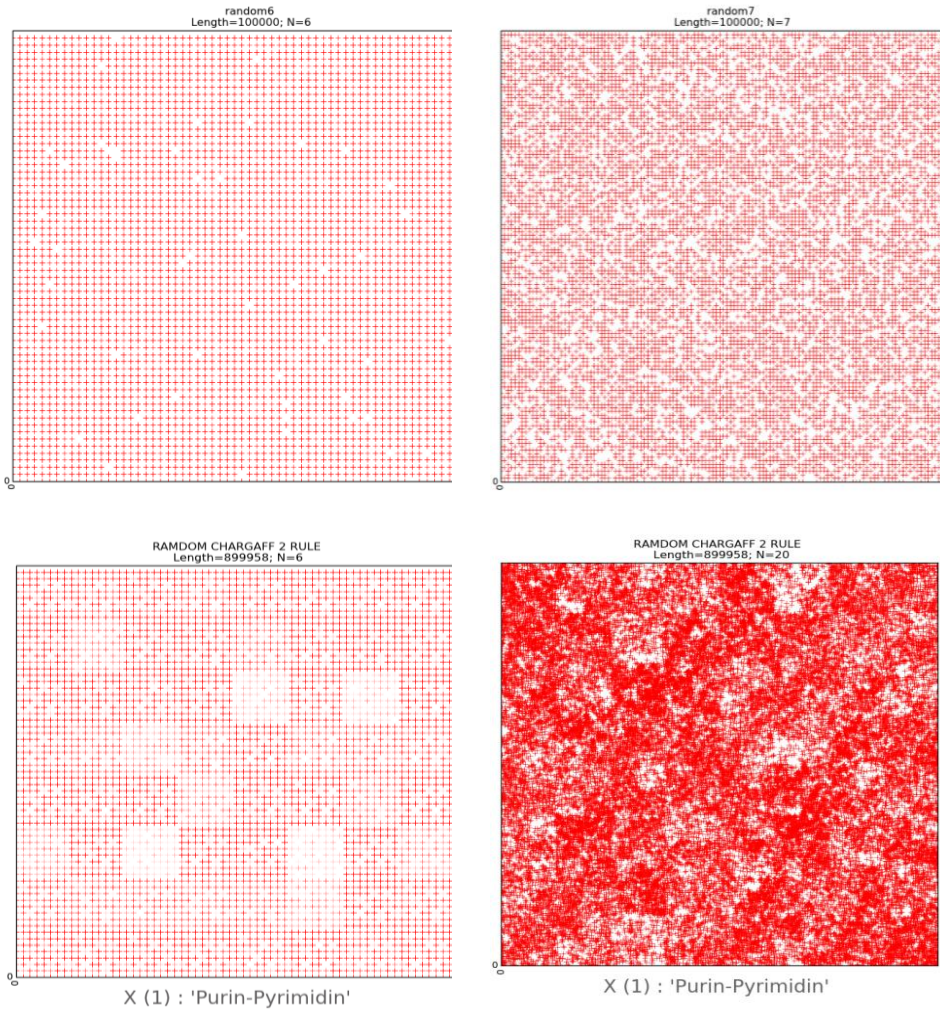


Рис.11. Верхний ряд - иллюстрации двумерного представления состава случайно генерированной нуклеотидной последовательности без соблюдения правил Чаргаффа. Нижний ряд - примеры двумерного представления нуклеотидного состава случайно генерированной последовательности с учетом второго правила Чаргаффа. Осям абсцисс и ординат соответствуют десятичные представления двоичного кодирования каждого *N*-плета.

Таким образом, двумерная визуализация цепочек азотистых оснований позволяет отобразить варианты выполнения количественных правил Чаргаффа [1,11] с применением аппарата конечной геометрии [21]. Это обстоятельство может помочь в исследовании внутренних симметрий и других характеристик нуклеиновых кислот для изучения сложных взаимосвязей между живыми организмами.

Нами были построены визуальные представления ДНК различных видов пенициллина. Полученные результаты свидетельствуют о том, что геномы этой группы как правило, генерируют мозаики высокой плотности, напоминающие мозаики случайных последовательностей, что свидетельствует о высоком разнообразии нуклеотидного состава. Возможно, медицинское значение пенициллина связано именно с этой особенностью.

Таким образом, методы двумерной визуализации представляются полезными для изучения скрытых закономерностей в хромосомах, а также для классификации и сравнительного анализа различных геномов с возможными применениями в биотехнологии и медицине.

3.3 Визуализация нуклеиновых кислот в трех одномерных пространствах физико-химических параметров нуклеотидов

Как отмечалось, бинарные суб-алфавиты связаны между собой операцией сложения по модулю два и задают пространство со свойствами, при которых координаты каждой точки связаны между собой. В связи с этим, имеет смысл рассматривать каждое измерение в отдельности. Существует три параметрически одномерных связанных между собой пространства визуализации. Использование параметрически одномерных координатных осей {X}, {Y} и {Z} дает тройку различных отображений с использованием соответствующих суб-алфавитов. Ось абсцисс кодирует порядковый номер *N*-плета в последовательности, ось ординат кодирует упорядоченные по возрастанию десятичные значения двоичного представления каждого *N*-плета (примечание: сама визуализация двумерная, но параметрическое измерение одно).

На рис. 12 приведен пример визуализации хромосомы человека, на которой хорошо видны области с различным нуклеотидным составом. Эти специфические регионы отмечены на рисунке стрелками и могут быть визуализированы в различном масштабе в двумерных пространствах визуализации для их детального анализа.

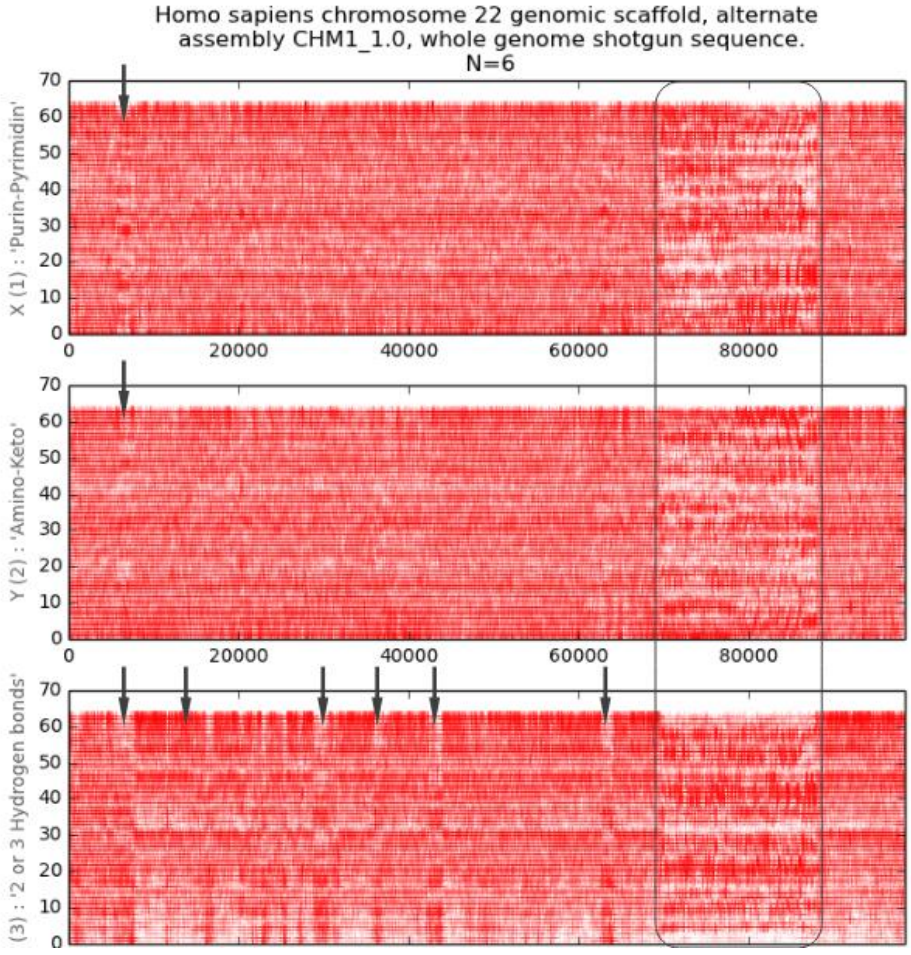


Рис.12. Визуализация трехканального представления нуклеотидного состава фрагмента 22-й хромосомы Homo Sapiens. Каждой из трех проекций соответствует двоично-оппозиционный суб-алфавит. В каждом канале ось абсцисс кодирует порядковый номер *N*-плета, ось ординат кодирует упорядоченные по возрастанию десятичные значения двоичного представления *N*-плетов. Стрелками выделены некоторые области с различным нуклеотидным составом. Большая область с отличающимся нуклеотидным составом обведена. Видно, что в различных частях хромосомы нуклеотидный состав может отличаться по каждому из каналов.

На рис. 13 и 14 к каждому из трех суб-алфавитов дополнительно приведена интегрально-одномерная визуализация суммарного числа единиц в кодах *N*-меров. Полученные графики позволяют оценить изменения в нуклеотидном составе при прочтении фрагмента молекулы от начала до конца. Глубина регистрируемых изменений определяется масштабирующим параметром *N*.

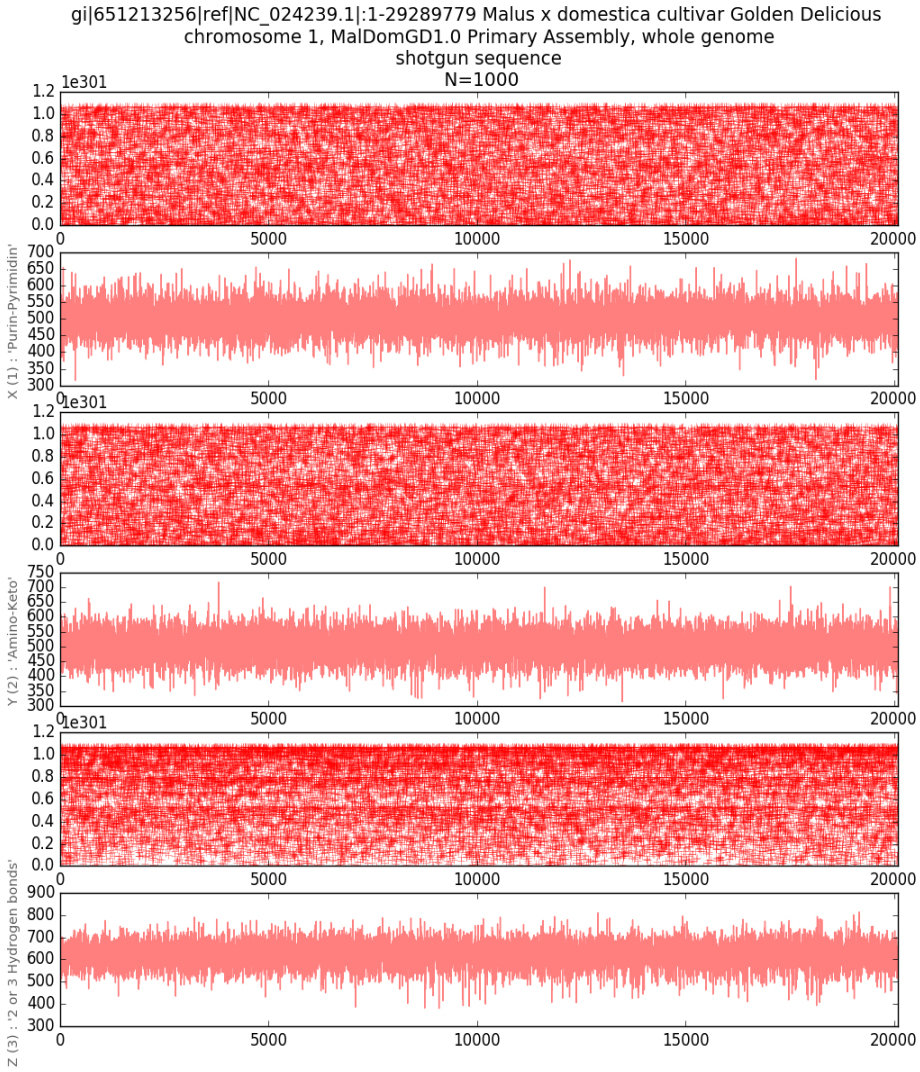


Рис.13. Визуализация трехканального представления нуклеотидного состава фрагмента 1-й хромосомы яблока. Каждому из трех рядов соответствует двоично-оппозиционный суб-алфавит. Ось абсцисс кодирует порядковый номер 1000-плета, ось ординат кодирует количество единиц в 1000-плете.

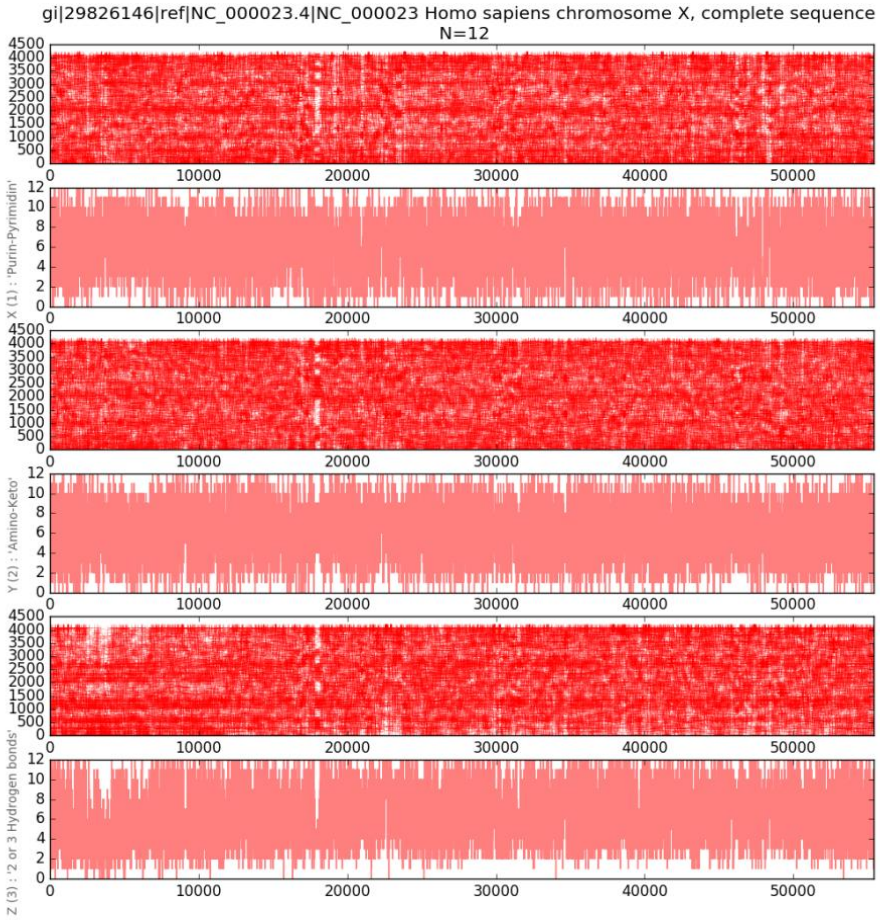


Рис.14. Визуализация трехканального представления нуклеотидного состава Х-хромосомы человека. Каждому из трех рядов соответствует двоично-оппозиционный суб-алфавит. Ось абсцисс кодирует порядковый номер 12-плета, ось ординат кодирует количество единиц в 12-плете.

Параметрически одномерные методы визуализации удобны тем, что позволяют отобразить нуклеотидный состав хромосомы, так, как его невозможно отобразить в двумерной и трехмерной проекциях. В связи с этим, описанные методы одномерной визуализации представляются информативными и перспективными для дальнейших исследований.

Отметим, что трёхканальное представление сочетается с классической теорией цветового восприятия (RGB), в которой считается, что глаз воспринимает три основных цвета: красный (red), зеленый (green) и синий (blue), а комбинации трёх основных цветов позволяют получить остальные цвета. Эта теория упоминается в [15] в связи с генетическими алгебрами. Каждый из трёх каналов одномерной визуализации можно сопоставить одному из трёх основных цветов. Интенсивность цвета каждой точки двумерной визуализации различная, поэтому двумерные и трехмерные представления позволяют учитывать сочетания цветов. Это позволяет усилить цветовое восприятие в генетике и открывает новые возможности для параметрической визуализации в соответствии с изложенным методом (однако, наши эксперименты показали, что это значительно увеличивает счётное время).

Для авторского метода параметрической визуализации предлагается введение нового термина «генетическая геометрия» или «генометрия» как основа соответствующего научного направления в области молекулярно-биологической биосемиотики [19].

4. Выводы

Результатом исследования является достижение поставленной цели по развитию методов визуализации длинных нуклеотидных последовательностей. Продемонстрированы связи молекулярно-генетических систем с двоичной системой счисления и матрицами Адамара. Гипотеза о возможности визуализации внутренних симметрий в нуклеотидном составе подтвердилась. Нуклеиновые кислоты имеют наглядное представление. Параметрическая визуализация как фрагментов, так и целых молекул ДНК и РНК позволила обосновать их связь с геометрическими мозаиками различных типов (см. например рис. 4-9). Предложенный метод позволяет оценить виды соотношений между присутствующими и отсутствующими *N*-мерами в ДНК различных организмов (этим соотношениям может быть свойственна фрактально-кластерная организация, наглядный пример — рис. 9). Масштабирующий параметр *N* позволяет исследовать геном на множестве уровней детализации для поиска скрытых симметрий и закономерностей.

Появление обоснованных методов сопоставления геометрических представлений генотипов с теми или иными фенотипическими признаками расширяет методы исследований в области молекулярной генетики. Кроме того, открывается возможность моделирования псевдослучайных нуклеотидных последовательностей с соблюдением феноменологических правил Чаргаффа для их визуализации и дальнейших исследований. Масштабно-параметрическая визуализация нуклеотидного состава способствует углубленному пониманию генетических явлений не только за счет упрощения восприятия, но также и за счет применения адаптивных нейросетевых технологий, поскольку структура хромосом живых организмов, представленная в двоичном коде, соответствует формату бинарных искусственных нейронных сетей [13].

Авторский метод визуализации является дополнительным критерием классификации и выявления межвидовых взаимосвязей. В связи с этим, современные онтологии и тезаурусы для организации и хранения молекулярно-генетических данных могут быть оснащены вариантами визуализации для образовательных целей, а также для представления и поиска биологической информации. Предложенный метод также может помочь продвинуться в понимании принципов функционирования иммунной системы при распознавании нуклеотидного состава вирусов, ДНК паразитов, а также в пищевых цепях и экосистемах. Геометрические представления могут помочь в изучении механизма точечных мутаций и систем CRISPR-Cas [16]. Это становится возможным за счет наглядной интерпретации базовых характеристик фрагментов полинуклеотидов определенного нуклеотидного состава с визуализацией конечной геометрии и структуры генетического кода.

Изложенные результаты позволяют говорить об авторских методах визуализации нуклеиновых кислот как о масштабно-параметрической модели ДНК, дополняющей структурную модель двойной спирали Дж. Уотсона и Ф. Крика [17,18].

Автор выражает благодарность Сергею Валентиновичу Петухову, Виталию Ивановичу Свирину, Константину Владимировичу Плешакову, Денису Сергеевичу Изюмову и Дмитрию Витальевичу Салонину за плодотворные научные дискуссии.

Список использованных источников

1. Chargaff E, Lipshitz R, Green C (1952)."Composition of the deoxypentose nucleic acids of four genera of sea-urchin" (PDF). J Biol Chem.195 (1): 155–160. PMID 1493836
2. S.V.Petoukhov, M.He. Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications. 2010, Hershey, USA: IGI Global. 271 p.
3. N.A.Balonin, Y.N.Balonin, D.Z. Djokovic, D.A. Karbovskiy, M.B.Sergeev. Construction of symmetric Hadamard matrices https://arxiv.org/abs/1708.05098

4. Georgiou, S.; Koukouvinos, C.; Seberry, J. (2003). "Hadamard matrices, orthogonal designs and construction algorithms".Designs 2002: Further computational and constructive design theory. Boston: Kluwer. pp. 133–205. ISBN 1-4020-7599-5.
5. I.V. Stepanian, S.V. Petoukhov. The matrix method of representation, analysis and classification of long genetic sequences http://arxiv.org/pdf/1310.8469.pdf
6. Х., Хармут Применение методов теории информации в физике / Х. Хармут. - М.: Мир, 2016. - 344 с.
7. Ferleger, Sergei V. (March 1998). RUC-Systems In Non-Commutative Symmetric Spaces (Technical report). MP-ARC-98-188.
8. Jeffrey H.J. (1990). Chaos game representation of gene structure. - Nucleic Acids Research, Vol.18, No.8, p. 2163-2170.
9. Feldman, David P. (2012), "17.4 The chaos game", Chaos and Fractals: An Elementary Introduction, Oxford University Press, pp. 178–180, ISBN 9780199566440.
10. G. Darvas, A.A. Koblyakov, S.V.Petoukhov, I.V.Stepanyan. Symmetries in molecular-genetic systems and musical harmony // Symmetry: Culture and Science Vol. 23, No. 3-4, 343-375, 2012 http://symmetry.hu/scs_online/SCS_23_3-4.pdf
11. Rudner, R; Karkas, JD; Chargaff, E (1968)."Separation of B. SubtilisDNA into complementary strands. 3. Direct analysis".Proceedings of the National Academy of Sciences of the United States of America. 60(3): 921–2. doi:10.1073/pnas.60.3.921. PMC 225140. PMID 4970114.
12. Townsend JP, Su Z, Tekle Y (2012). "Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny". Genetics. 61(5): 835–849. doi:10.1093/sysbio/sys036. PMID 22389443.
13. Степанян И.В., Зиеп Н.Н. Растущие сверточные нейроподобные структуры для задач распознавания статических образов // Нейрокомпьютеры: разработка, применение. 2018. № 5. С. 4-11.
14. ftp://ftp.ncbi.nlm.nih.gov/
15. Петухов, С.В. Матричная генетика, алгебры генетического кода, помехоустойчивость. — М.: РХД. — 2008.
16. Ikeda T., Tanaka W., Mikami M., Endo M., Hirano H.-Y. Generation of artificial drooping leaf mutants by CRISPR-Cas9 technology in rice // Genes & Genetic Systems. — 2016. — Vol. 90, no. 4. — P. 231—235. — DOI:10.1266/ggs.15-00030. — PMID 26617267.
17. Crick FH, Wang JC, Bauer WR (April 1979). "Is DNA really a double helix?" (PDF). J. Mol. Biol. 129 (3): 449–57. doi:10.1016/0022-2836(79)90506-0. PMID 458852.
18. Wilkins MH, Stokes AR, Wilson HR (1953). "Molecular Structure of Deoxypentose Nucleic Acids" (PDF). Nature. 171 (4356):738–740. Bibcode: 1953 Natur. 171..738W. doi:10.1038/171738a0. PMID 13054693.
19. Sharov A. (1992). Biosemiotics: functional-evolutionary approach to the analysis of the sense of information. In: Biosemiotics: The Semiotic Web 1991. T.A.Sebeok and J.Umiker-Sebeok (eds.), 345-373. Berlin: Mouton de Gruyter.
20. Waterman M.S. Introduction to Computational Biology. Map, Sequences and Genomes. London: Chapman & Hall, 1995. xvi + 432 pp.
21. Batten, Lynn Margaret (1997), Combinatorics of Finite Geometries, Cambridge University Press, ISBN 0521590140
22. Petoukhov S.V., Petukhova E.S., Svirin V.I. New Symmetries and Fractal-Like Structures in the Genetic Coding System. – Advances in Intelligent Systems and Computing, v. 754, 2018, p. 588-600, https://doi.org/10.1007/978-3-319-91008-6_60
23. Mcdonnell K, Waters N, Howley E, Abram F. Chordomics: a visualisation tool for linking function to phylogeny in microbiomes. Bioinformatics. 2019;
24. Mathema VB, Dondorp AM, Imwong M. OSTRFPD: Multifunctional Tool for Genome-Wide Short Tandem Repeat Analysis for DNA, Transcripts, and Amino Acid Sequences with Integrated Primer Designer. Evol Bioinform Online. 2019;15:1176934319843130.

25. Iacoangeli A, Al khleifat A, Sproviero W, et al. DNAscan: personal computer compatible NGS analysis, annotation and visualisation. BMC Bioinformatics. 2019;20(1):213.
26. Martens KJA, Van beljouw SPB, Van der els S, et al. Visualisation of dCas9 target search in vivo using an open-microscopy framework. Nat Commun. 2019;10(1):3552.

Scientific Visualization, 2020, volume 12, number 3, pages 61 - 78, DOI: 10.26583/sv.12.3.06

A multiscale model of nucleic acid imaging

I.V. Stepanyan[1]

Institute of Machine Science named after A.A.Blagonravov of the RAS

[1] ORCID: 0000-0003-3176-5279, neurocomp.pro@gmail.com

Abstract

The paper describes new results in the field of algebraic biology, where matrix methods are used [Petukhov, 2008, 2012, 2013; Petuhov, He, 2010] with the transition from matrix algebra to discrete geometry and computer visualization of the genetic code. The algorithms allow to display the composition of sequences of nitrogenous bases in parametric spaces of various dimensions. Examples of visualization of the nucleotide composition of genetic sequences of various species of living organisms are given. The analysis was carried out in the spaces of binary orthogonal Walsh functions taking into account the physical and chemical parameters of the nitrogen bases. The results are compared with the rules of Erwin Chargaff concerning genetic sequences in the composition of DNA molecules. The developed method makes it possible to substantiate the relationship between DNA and RNA molecules with fractal and other geometric mosaics, reveals the orderliness and symmetries of polynucleotide chains of nitrogen bases and the noise immunity of their visual representations in the orthogonal coordinate system. The proposed methods can serve to simplify the researchers' perception of long chains of nitrogenous bases through their geometrical visualization in parametric spaces of various dimensions, and also serve as an additional criterion for classifying and identifying interspecific relationships.

Keywords: visualization algorithms, Walsh functions, Chargaff's rules, multidimensional analysis, nucleotide composition, fractals, bioinformatics, DNA, chromosomes, symmetries.

References

1. Chargaff E, Lipshitz R, Green C (1952). "Composition of the deoxypentose nucleic acids of four genera of sea-urchin" (PDF). J Biol Chem.195 (1): 155–160. PMID 1493836
2. S.V.Petoukhov, M.He. Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications. 2010, Hershey, USA: IGI Global. 271 p.
3. N.A.Balonin, Y.N.Balonin, D.Z. Djokovic, D.A. Karbovskiy, M.B.Sergeev. Construction of symmetric Hadamard matrices https://arxiv.org/abs/1708.05098
4. Georgiou, S.; Koukouvinos, C.; Seberry, J. (2003). "Hadamard matrices, orthogonal designs and construction algorithms".Designs 2002: Further computational and constructive design theory. Boston: Kluwer. pp. 133–205. ISBN 1-4020-7599-5.
5. I.V. Stepanian, S.V. Petoukhov. The matrix method of representation, analysis and classification of long genetic sequences http://arxiv.org/pdf/1310.8469.pdf
6. H., Harmuth Applying of methods of theory of information in phisics / - Moscow.: Mir, 2016. - p. 344.
7. Ferleger, Sergei V. (March 1998). RUC-Systems In Non-Commutative Symmetric Spaces (Technical report). MP-ARC-98-188.
8. Jeffrey H.J. (1990). Chaos game representation of gene structure. - Nucleic Acids Research, Vol.18, No.8, p. 2163-2170.
9. Feldman, David P. (2012), "17.4 The chaos game", Chaos and Fractals: An Elementary Introduction, Oxford University Press, pp. 178–180, ISBN 9780199566440.

10. G. Darvas, A.A. Koblyakov, S.V.Petoukhov, I.V.Stepanyan. Symmetries in molecular-genetic systems and musical harmony // Symmetry: Culture and Science Vol. 23, No. 3-4, 343-375, 2012 http://symmetry.hu/scs_online/SCS_23_3-4.pdf
11. Rudner, R; Karkas, JD; Chargaff, E (1968). "Separation of B. SubtilisDNA into complementary strands. 3. Direct analysis".Proceedings of the National Academy of Sciences of the United States of America. 60(3): 921–2. doi:10.1073/pnas.60.3.921. PMC 225140. PMID 4970114.
12. Townsend JP, Su Z, Tekle Y (2012). "Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny". Genetics. 61(5): 835–849. doi:10.1093/sysbio/sys036. PMID 22389443.
13. Stepanyan I.V., Ziep N.N. Growing convolutional neural-like structures for problems of recognition of static images // Neurocomputers: development, application. 2018. № 5. pp. 4-11.
14. ftp://ftp.ncbi.nlm.nih.gov/
15. Petukhov, S.V. Matrix genetics, algebra of genetic code, noise immunity. - M.: RHD. - 2008.
16. Ikeda T., Tanaka W., Mikami M., Endo M., Hirano H.-Y. Generation of artificial drooping leaf mutants by CRISPR-Cas9 technology in rice // Genes & Genetic Systems. — 2016. — Vol. 90, no. 4. — P. 231—235. — DOI:10.1266/ggs.15-00030. — PMID 26617267.
17. Crick FH, Wang JC, Bauer WR (April 1979). "Is DNA really a double helix?" (PDF). J. Mol. Biol. 129 (3): 449–57. doi:10.1016/0022-2836(79)90506-0. PMID 458852.
18. Wilkins MH, Stokes AR, Wilson HR (1953). "Molecular Structure of Deoxypentose Nucleic Acids" (PDF). Nature. 171 (4356):738–740. Bibcode: 1953 Natur. 171..738W. doi:10.1038/171738a0. PMID 13054693.
19. Sharov A. (1992). Biosemiotics: functional-evolutionary approach to the analysis of the sense of information. In: Biosemiotics: The Semiotic Web 1991. T.A.Sebeok and J.Umiker-Sebeok (eds.), 345-373. Berlin: Mouton de Gruyter.
20. Waterman M.S. Introduction to Computational Biology. Map, Sequences and Genomes. London: Chapman & Hall, 1995. xvi + 432 pp.
21. Batten, Lynn Margaret (1997), Combinatorics of Finite Geometries, Cambridge University Press, ISBN 0521590140
22. Petoukhov S.V., Petukhova E.S., Svirin V.I. New Symmetries and Fractal-Like Structures in the Genetic Coding System. – Advances in Intelligent Systems and Computing, v. 754, 2018, p. 588-600, https://doi.org/10.1007/978-3-319-91008-6_60
23. Mcdonnell K, Waters N, Howley E, Abram F. Chordomics: a visualisation tool for linking function to phylogeny in microbiomes. Bioinformatics. 2019;
24. Mathema VB, Dondorp AM, Imwong M. OSTRFPD: Multifunctional Tool for Genome-Wide Short Tandem Repeat Analysis for DNA, Transcripts, and Amino Acid Sequences with Integrated Primer Designer. Evol Bioinform Online. 2019;15:1176934319843130.
25. Iacoangeli A, Al khleifat A, Sproviero W, et al. DNAscan: personal computer compatible NGS analysis, annotation and visualisation. BMC Bioinformatics. 2019;20(1):213.
26. Martens KJA, Van beljouw SPB, Van der els S, et al. Visualisation of dCas9 target search in vivo using an open-microscopy framework. Nat Commun. 2019;10(1):3552.