БИОФИЗИКА, 2024, том 69, № 2, с. 246–256

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 57.015 + 573.2

РАСПРЕДЕЛЕНИЕ ЗНАЧЕНИЙ GC-СОСТАВА ФРАГМЕНТОВ В ПРОСТРАНСТВЕННОЙ СТРУКТУРЕ ГЕНОМОВ ХЛОРОПЛАСТОВ, БАКТЕРИЙ И МИТОХОНДРИЙ

© 2024 г. М.Ю. Сенашова*,#, М.Г. Садовский*, **, *,##**

**Институт вычислительного моделирования СО РАН, Академгородок, 50/44, Красноярск, 660036, Россия*

#E-mail: msen@icm.krasn.ru

***Федеральный Сибирский научно-клинический центр ФМБА России, Коломенская ул., 26, Красноярск, 660037, Россия*

****Сибирский федеральный университет, Свободный просп., 79, Красноярск, 660041, Россия*

##E-mail: msad@icm.krasn.ru

Поступила в редакцию 03.01.2023 г.
После доработки 26.02.2024 г.
Принята к публикации 06.03.2024 г.

Рассмотрено пространственное распределение значений GC-состава фрагментов геномов хлоропластов, митохондрий и бактерий. Обнаружено, что значения GC-состава фрагментов для большинства геномов распределены не хаотически, а упорядоченно. Были обнаружены два основных типа распределения – градиентное и центрально-симметричное. У геномов хлоропластов встречается только градиентное распределение. У бактерий для GC-бедных геномов наблюдается центрально-симметричное распределение, а для GC-богатых – градиентное. У митохондрий встречаются оба варианта распределений, причем тип распределения зависит от вида организма.

Ключевые слова: GC-состав, частотные словари, триплеты, структура, пространственное распределение.

DOI: 10.31857/S0006302924020049, **EDN:** OVJRZZ

Изучение особенностей и деталей структуры нуклеотидных последовательностей является важнейшей задачей биологии в настоящее время. Исследования ведутся в двух аспектах — структурно-функциональном и эволюционном. Выявление связи между структурными компонентами и соответствующими им функциями представляет собой классическую проблему молекулярной и системной биологии, и, несмотря на обширный поток публикаций и исследований в этом направлении, она все еще далека от завершения. Более того, исследователи выявляют все новые и новые структурные элементы либо новые виды и формы взаимодействий и взаимоотношений между структурными элементами биологических макромолекул, этому способствует развитие техники и инструментов исследований.

Понятна важность таких исследований с точки зрения эволюционных процессов. Изучение особенностей структуры биологических макромолекул у разных организмов позволяет составить более точную картину эволюции тех или иных биологических систем – от вполне конкретных видов до экосистем и глобальных сообществ.

Кроме того, затруднения в исследованиях такого рода всегда вызывают выбор и качество того биологического материала, который берется в рассмотрение. Дело даже не в ошибках секвенирования и/или аннотирования генетических последовательностей, неизбежных во многих случаях, а в большой сложности таких объектов, как геномы либо отдельные хромосомы. Рассматривая эти объекты, приходится анализировать набор характеристик – структуру, функцию и филогению. Эти характеристики очень сильно взаимодействуют и сильно влияют друг на друга. Причем это влияние далеко не всегда удается выделить в качестве отдельного и независимого фактора.

Прокариотические организмы с этой точки зрения являются более удобными объектами для исследования, чем эукариотические; геном бактерий заметно короче генома эукариот и всегда представлен одной хромосомой. Однако еще удобнее использовать геномы органелл, в нашем случае – митохондрий и хлоропластов, поскольку

для них полностью исключается влияние различий в кодируемых функциях: в пределах одной группы органелл функциональные различия отсутствуют.

Такой параметр, как GC-состав очень часто используется в исследованиях структуры геномов и функций отдельных участков. Под GC-составом понимают долю гуанина и цитозина по отношению к общему числу нуклеотидов в рассматриваемой нуклеотидной последовательности. Пары «гуанин–цитозин» и «аденин–тимин» отличаются количеством водородных связей между ними. Это обуславливает разные физические свойства у GC-бедных и GC-богатых нуклеотидных последовательностей. Использованию GC-состава в исследованиях структуры геномов и функций отдельных участков, именно с точки зрения физических свойств, посвящено много работ, как в случае геномов хлоропластов [1–8] и геномов бактерий [9–15], так и митохондрий [16–22]. В данной работе нуклеотидные последовательности рассматриваются с точки зрения биоинформатики. Изучается распределение значений GC-состава выделенных фрагментов геномов хлоропластов, митохондрий и бактерий в пространственной структуре генома, полученной на основе частотных словарей этих фрагментов. Возникает вопрос – имеется ли упорядоченность значений GC-состава в нуклеотидных последовательностях либо они расположены случайным образом. Естественно было бы ожидать случайное распределение значений GC-состава, но, как показывают результаты работы, наблюдается достаточно высокая упорядоченность такого распределения.

МАТЕРИАЛЫ И МЕТОДЫ

Введем основные понятия. Прежде чем перейти к изложению результатов, поясним смысл используемого ниже термина «пространство». Нуклеотидная последовательность является линейным объектом и в таковом качестве является одномерной. При этом в физическом трехмерном (евклидовом) пространстве у нуклеотидной последовательности могут наблюдаться и двумерные и даже трехмерные структуры: типичный пример – спиральная структура ДНК, а также плотная «намотка» нуклеотидной последовательности на гистон. Однако эти понятия пространства не будут использоваться в дальнейшей работе.

Мы будем рассматривать генетическую последовательность длины L, состоящую из символов алфавита $\aleph = \{A,C,G,T\}$. Для этой последовательности будем составлять частотный словарь толщины 3. Частотный словарь W_3 толщины 3 символьной последовательности, соответствующей ДНК – это список всех троек $\nu_1\nu_2\nu_3$ идущих подряд нуклеотидов с указанием частот этих троек; всего может быть 64 триплета. Заметим, что приведенное определение частотного словаря триплетов является частным случаем: при подсчете числа триплетов окно считывания может перемещаться на один, два, три, четыре и вообще произвольное число нуклеотидов, тем самым порождая разные частотные словари. Как правило, используется частотный словарь вида W_3^k, где $k = 1$.

Как и в работах А.Н. Горбаня с соавторами [10, 11], мы используем частотный словарь вида W_3^3, тем самым триплеты в нашей работе подсчитывались таким образом, что они полностью покрывают последовательность и при этом не пересекаются. Частота f_ω – это отношение числа копий n_ω данного триплета к общему числу всех триплетов N, где N – сумма всех n_ω:

$$f_\omega = \frac{n_\omega}{N}. \quad (1)$$

Всякий частотный словарь W_3^3 отображает геном в 64-мерное метрическое (евклидово) пространство. Иначе говоря, каждому фрагменту генома, для которого подсчитывается частотный словарь, ставится в соответствие точка в 64-мерном пространстве. Осям этого пространства соответствуют триплеты. Координатами точки являются значения частот триплетов. В результате разбиения генома на фрагменты он отображается во множество точек в пространстве триплетов.

Рассмотрим более детально предварительную обработку, которая ставит в соответствие данной генетической последовательности множество точек в 64-мерном пространстве триплетов. Делалось это следующим образом: последовательность сканировалась окном длины Δ с шагом t. Для каждого положения i-го окна определяли участок генетической последовательности, для которого вычисляли частотный словарь $W_3^{(i)}$ соответствующий i-й точке в 64-мерном пространстве. Кроме того, с каждой точкой в 64-мерном пространстве связывали номер центрального символа рассматриваемого участка, который совпадает с номером этого символа в последовательности. Данные для исследования брали в базе EMBL-банка. Для всех генетических последовательностей длина окна $\Delta = 603$, шаг $t = 11$.

В каждом из полученных частотных словарей один из 64 триплетов исключался, поскольку сумма всех частот в словаре равна 1, что порождает линейную связь, которая может давать ложный сигнал при последующей обработке (корреляци-

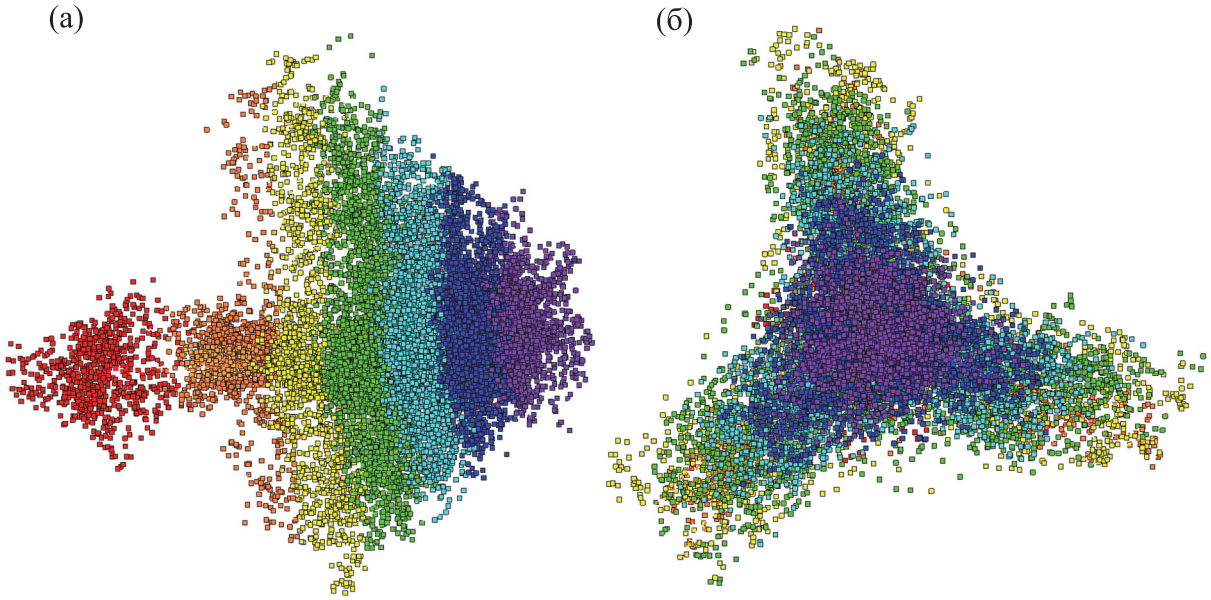

Рис. 1. Пространственное расположение GC-состава фрагментов генома *Anthoceros angustus* в пространстве 1-й и 2-й главных компонент (рис. 1а) и 2-й и 3-й главных компонент (рис. 1б).

онном анализе, определении главных компонент и т.п.).

В нашем случае целесообразнее исключать тот триплет, для которого стандартное отклонение, наблюдаемое по ансамблю всех частотных словарей, является минимальным: такой триплет дает наименьший вклад в различимость объектов (в предельном случае, когда стандартное отклонение равно 0, различий по этому триплету вовсе нет). Таким образом, рассматриваемое нами пространство точек становится 63-мерным. Для каждого генома исключаемый триплет будет своим.

Для визуализации полученного множества точек в программе VidaExpert (http://bioinfo-out.curie.fr/projects/vidaexpert/) строили проекцию данных в пространстве первых трех главных компонент, вычисленных для данного 63-мерного пространства. Тем самым мы сокращаем 63-мерное пространство до трехмерного; данные становятся обозримыми. При обработке геномов было обнаружено, что точки располагаются достаточно однотипно в проекции на плоскость, определяемую 1-й и 2-й главными компонентами. Для более полной картины рассматривалась также плоскость, определяемая 2-й и 3-й главными компонентами.

Идейно наша работа опирается на аналогичные исследования геномов хлоропластов, выполненные в работе [20]. В ней рассматривались геномы хлоропластов наземных растений; далее изучалось распределение точек, определяемых фрагментами геномов в пространстве частот триплетов, а в качестве преобладающей структуры такого распределения была выявлена трехлучевая с хвостом структура.

В нашей работе, кроме хлоропластов, объектом исследования являются геномы бактерий и геномы митохондрии животных и растений. Геномы бактерий значительно длиннее, а геномы митохондрий значительно короче геномов хлоропластов, что, по-видимому, может являться причиной найденных различий. Для всех геномов, кроме геномов митохондрий наземных растений и печеночных мхов, наблюдается трехлучевая структура. Под такой структурой мы понимаем распределение точек, представляющих частотные словари фрагментов в виде трех эллипсоидов (со значительным эксцентриситетом), пересекающихся одним из своих концов. В месте пересечения концентрируются точки, соответствующие некодирующим областям генома. При этом большие оси обнаруженных эллипсоидов отстоят на равные углы. Отметим, что для геномов митохондрий наземных растений и печеночных мхов все три эллипсоида совпадают и образуют единый эллипсоид.

ПРОСТРАНСТВЕННАЯ СТРУКТУРА GC-СОСТАВА ФРАГМЕНТОВ ГЕНОМОВ ХЛОРОПЛАСТОВ

Было рассмотрено 570 геномов хлоропластов из базы EMBL. Как было показано в работе [23], подавляющее большинство геномов хлоропластов наземных растений имеет трехлучевую

структуру с хвостом. Трехлучевая структура представлена на рис. 1б (проекция на плоскость второй и третьей главных компонент), кластер, образующий так называемый «хвост» виден на рис. 1а слева в плоскости первой и второй главных компонент. На рис 1а,б показана структура генома хлоропласта с раскрашенным распределением величины GC-состава на примере генома *Anthoceros angustus* (идентификатор AB086179 в EMBL). Исключен триплет GCG. Для всех геномов здесь и далее интервал имеющихся значений GC-состава разбивали на 7 подинтервалов равной длины. Интервалу с минимальными значениями соответствует фиолетовый цвет, с максимальными – красный. Было обнаружено, что распределение значений GC-состава фрагментов геномов хлоропластов по пространственной структуре однотипно. GC-состав фрагментов распределен по градиенту вдоль оси симметрии пространственной структуры генома – от меньших значений к большим, причем минимальные значения находятся в вершине трехлучевой структуры, а максимальные – в отдельно расположенном кластере, так называемом «хвосте» (рис. 1а).

ПРОСТРАНСТВЕННАЯ СТРУКТУРА GC-СОСТАВА ФРАГМЕНТОВ ГЕНОМОВ БАКТЕРИЙ

Были проанализированы 68 геномов бактерий. В работе [10] было показано, что пространственная структура бактерий делится на четыре типа в зависимости от GC-состава. Все типы представляют собой различные комбинации взаимного расположения треугольников. Эти треугольники образованы точками, соответствующими фрагментам генома, имеющим одинаковые относительные фазы прямого либо обратного стренда. Относительная фаза определяется положением стартового нуклеотида выделяемого фрагмента по отношению к началу кодирующего участка. При этом остаток от деления на 3 длины между началом кодирующего участка и выделенного фрагмента и определяет номер фазы. Значит, вершины одного треугольника содержат точки с одинаковыми относительными фазами одного стренда (прямого или обратного). Один треугольник содержит точки, относящиеся к прямому (обратному) стренду. При этом взаимное расположение вершин этих треугольников в пространстве первых трех главных компонент существенно зависит от GC-состава генома. Первый тип – «параллельные треугольники» – характерен для AT-богатых геномов (GC-состав имеет значения от примерно 25 до 35%). Затем идут «перпендикулярные треугольники» для которых GC-состав изменяется примерно от 35 до 50%. Постепенно происходит переход к вырожденному случаю (треугольники совпадают) в интервале значений примерно от 50 до 60%. И, наконец, при значениях GC-состава больше 60% мы имеем шестилучевую структуру, образованную повернутыми относительно друг друга треугольниками, расположенными в одной плоскости. Рассмотрим, как представлены значения GC-состава для всех четырех типов пространственной структуры геномов бактерий (рис. 2). На рисунках представлены геномы следующих бактерий: *Clostridium baratii str. Sullivan* (идентификатор CP006905 в EMBL), *Atopobium parvulum DSM 20469* (идентификатор CP001721 в EMBL), *Synechococcus* sp. *CC9311* (идентификатор CP000435 в EMBL), *Synechococcus* sp. *JA-3-3Ab* (идентификатор CP000239 в EMBL). Были исключены триплеты GCG, GGG, TAC и TTA соответственно. Как видно из рис. 2а,б, для GC-бедных геномов (значение GC-состава по геному в целом меньше 50%) характерно центрально-симметричное распределение значений GC-состава фрагментов. Минимальные значения находятся в центре структуры, максимальные – по краям. При центрально-симметричном распределении нет такой ярко выраженной градации по величине значений, как при градиентном, однако минимальные значения преимущественно располагаются в центре структуры, а максимальные по краям. Промежуточные значения в большой степени перемешаны между собой, но тем не менее можно заметить, что большее количество точек, относящееся к определенному интервалу значений, расположено дальше от центра структуры, чем большинство точек предыдущего интервала. Для GC-богатых геномов (значение GC-состава по геному в целом больше 50%) наблюдается градиентное распределение значений GC-состава фрагментов вдоль оси симметрии структуры геномов (рис. 2в,г).

ПРОСТРАНСТВЕННАЯ СТРУКТУРА GC-СОСТАВА ФРАГМЕНТОВ ГЕНОМОВ МИТОХОНДРИЙ

Было отобрано 418 митохондрий растений и животных, в среднем по 25 видов на тип, класс или группу. Пространственная структура геномов митохондрий имеет преимущественно трехлучевую структуру (пересечение трех эллипсоидов). Однако геномы митохондрий наземных растений и печеночных мхов имеют структуру, близкую к единственному эллипсоиду. Рассмотрим группы геномов, для которых наблюдается однотипное распределение значений GC-состава по пространственной структуре геномов.

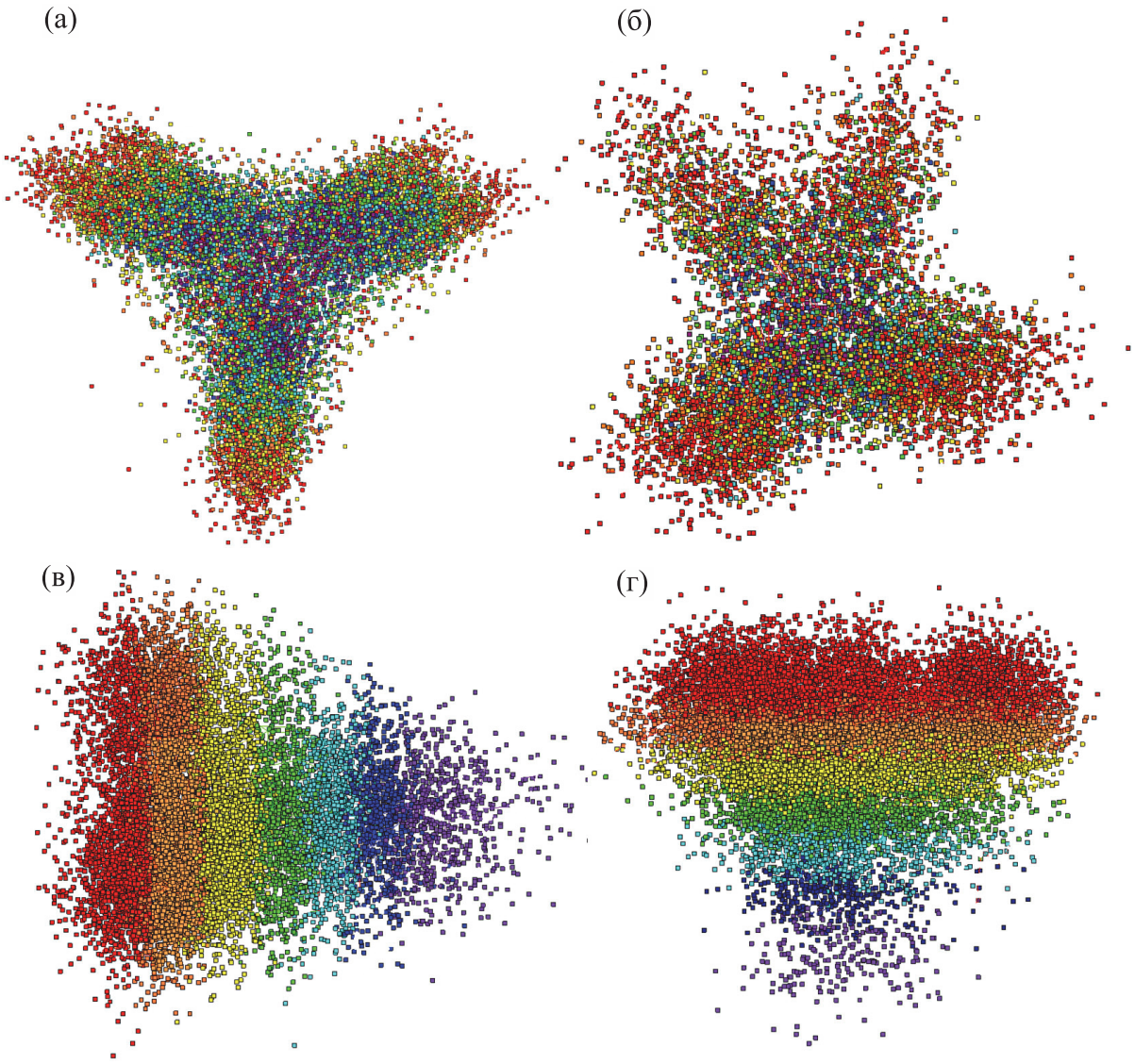

Рис. 2. Пространственное расположение GC-состава фрагментов генома *Clostridium baratii str. Sullivan* (рис. 2а), *Atopobium parvulum DSM 20469* (рис. 2б), *Synechococcus* sp. *CC9311* (рис. 2в) и *Synechococcus* sp. *JA-3-3Ab* (рис. 2г) в пространстве 1-й и 2-й главных компонент.

Рассмотрим сначала геномы митохондрий наземных растений и печеночных мхов как наиболее отличающихся ото всех остальных геномов митохондрий. Как уже было сказано выше, для наземных растений и печеночных мхов характерна структура геномов митохондрий эллипсоидной формы. Для их геномов характерно ярко выраженное градиентное распределение значений GC-состава вдоль оси симметрии структуры. На рис. 3 показано распределение значений GC-состава *Zea mays* (идентификатор AY506529 в EMBL, исключен триплет GCG). Как видно из рисунка, наблюдается градиентное распределение значений GC-состава фрагментов вдоль оси симметрии структуры генома. Аналогично выглядит структура генома и распределение GC-состава для печеночных мхов.

Для одноклеточных водорослей, обычных мхов и высших грибов характерно градиентное распределение значений GC-состава фрагментов вдоль оси симметрии структуры генома. На рис. 4 показано распределение значений GC-состава фрагментов на примере *Mesostigma viride* (идентификатор AF353999 в EMBL, исключен триплет *ccg*), *Physcomitrella patens* (идентификатор AB251495 в EMBL, исключен триплет *ccg*) и *Agaricus bisporus* var. *bisporus H97* (идентификатор JX271275 в EMBL, исключен триплет GCG).

Многоклеточные водоросли и низшие грибы не имеют ярко выраженного градиентного рас-

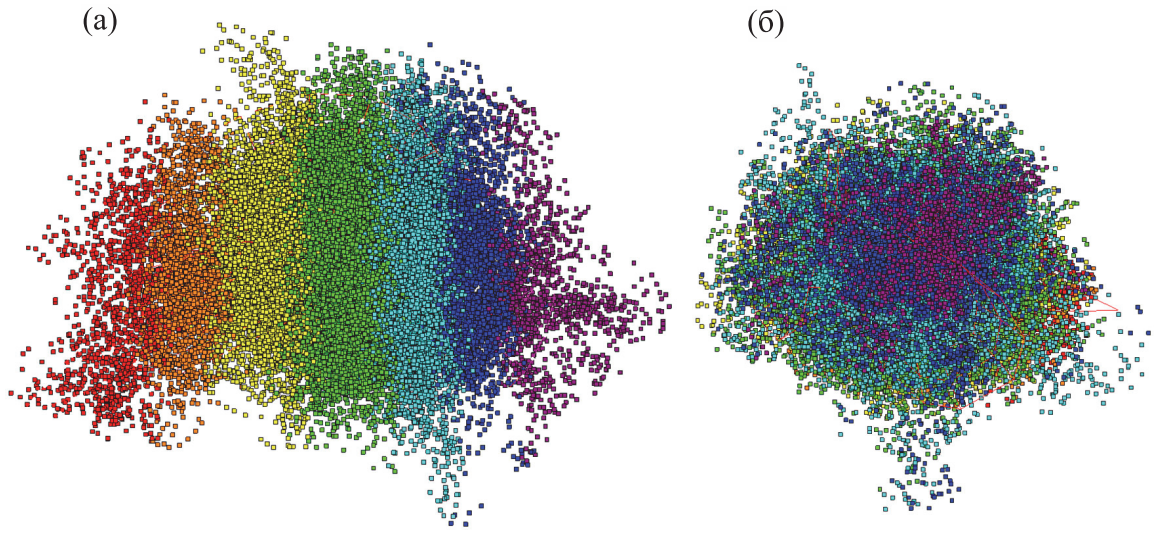

Рис. 3. Пространственное расположение GC-состава фрагментов генома *Zea mays* в пространстве 1-й и 2-й главных компонент (рис. 3а) и 2-й и 3-й главных компонент (рис. 3б), компонент (рис. 3а) и 2-й и 3-й главных компонент (рис. 3б).

пределения. Тем не менее можно заметить, что минимальные значения GC-состава располагаются большей частью на одном конце оси симметрии структуры, а максимальные – на другом. Распределение GC-состава фрагментов по структуре геномов показано на рис. 5 на примере *Saccharina japonica* (идентификатор AP011493 в EMBL, исключен триплет CGC) и *Candida viswanathii* (идентификатор EF536359 в EMBL, исключен триплет CCG).

Перейдем к рассмотрению геномов митохондрий животных.

Для геномов митохондрий насекомых, паукообразных и ракообразных характерно центральносимметричное распределение значений GC-состава фрагментов по пространственной структуре геномов. Минимальные значения GC-состава расположены в центре трехлучевой структуры геномов, максимальные по краям. Промежуточные значения не имеют ярко выраженной градации по значениям, но большинство точек предыдущего интервала значений находится ближе к центру, чем большинство точек следующего интервала. На рис. 6 показано центральносимметричное распределение значений GC-состава на примере *Homalodisca vitripennis* (идентификатор AY875213 в EMBL, исключен триплет CGG), *Liphistius erawan* (идентификатор JQ407803 в EMBL, исключен триплет CGG) и *Trigoniophthalmus alternatus* (идентификатор EU016193 в EMBL, исключен триплет GCG).

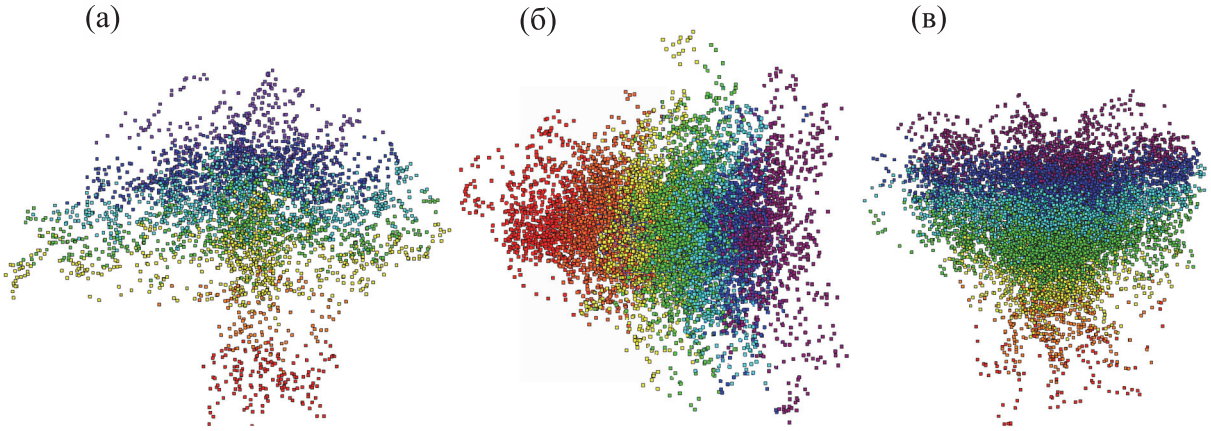

Рис. 4. Пространственное расположение GC-состава фрагментов генома *Mesostigma viride* в пространстве 1-й и 3-й главных компонент (рис. 4 а), *Physcomitrella patens* в пространстве 2-й и 3-й главных компонент (рис. 4 б) и *Agaricus bisporus* var. *bisporus H97* в пространстве 1-й и 3-й главных компонент (рис. 4 в).

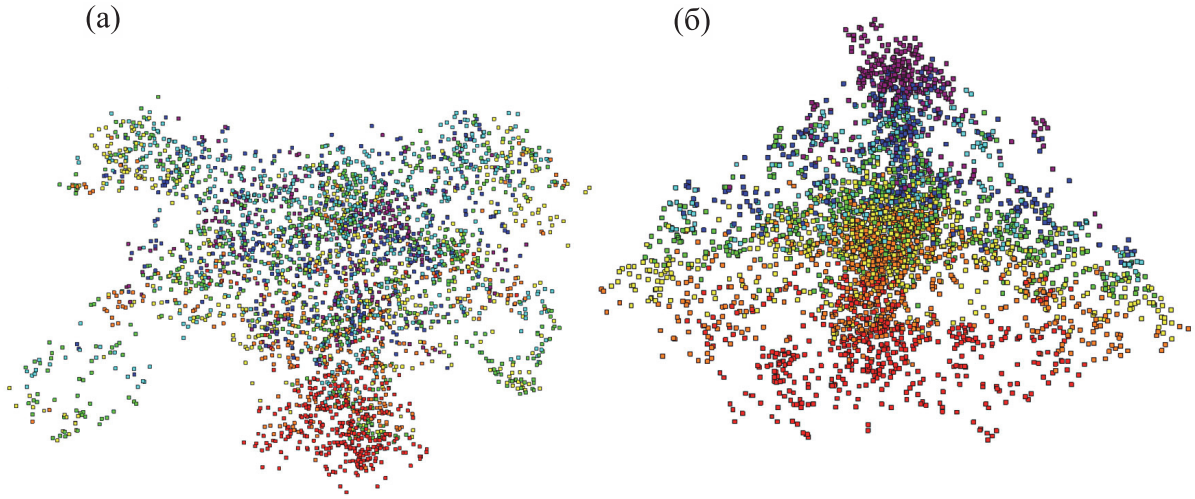

Рис. 5. Пространственное расположение GC-состава фрагментов генома *Saccharina japonica* (рис. 5 а) и *Candida viswanathii* (рис. 5б) в пространстве 1-й и 3-й главных компонент.

Следующую группу геномов составляют губки, моллюски, плоские, кольчатые и круглые черви. Для геномов этой группы встречаются распределение значений GC-состава в виде неявного градиентного распределения и центральносимметричное. На рис. 7 представлены распределения на примере *Callyspongia plicifera* (идентификатор EU237477 в EMBL, исключен триплет CGC) и *Echinococcus equinus* (идентификатор AF346403 в EMBL, исключен триплет *cgc*). Для позвоночных животных не удалось выявить какую-либо типичную картину распределения GC-состава фрагментов внутри классов и между классами.

УСТОЙЧИВОСТЬ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ GC-СОСТАВА ФРАГМЕНТОВ ГЕНОМОВ

Чтобы выяснить, насколько устойчивым является обнаруженное пространственное распределение GC-состава фрагментов геномов, были рассмотрены такие распределения для различной длины окна Δ. Распределения были рассмотрены в интервале сохранения тех структур геномов, которые представлены в статье. Для геномов хлоропластов и митохондрий этот интервал Δ равен от 300 до 1500, для бактерий – от 600 до 3000. Поскольку выше мы рассматривали распределение величины GC-состава фрагментов геномов для длины окна Δ = 600, то для выяснения устойчивости распределения были взяты длины окна Δ, равные 300, 900, 1200 и 1500 для хлоропластов и митохондрий, и окна 900, 1200, 1500 и 3000 – для бактерий. На рис. 8 показаны полученные распределения. В качестве примера для хлоропластов взят геном *Triticum aestivum* (ID AB042240), для бактерий – *Clostridium baratii* str. Sullivan (ID CP006905), а для митохондрий – *Cladonia petrophila* (ID MG941021).

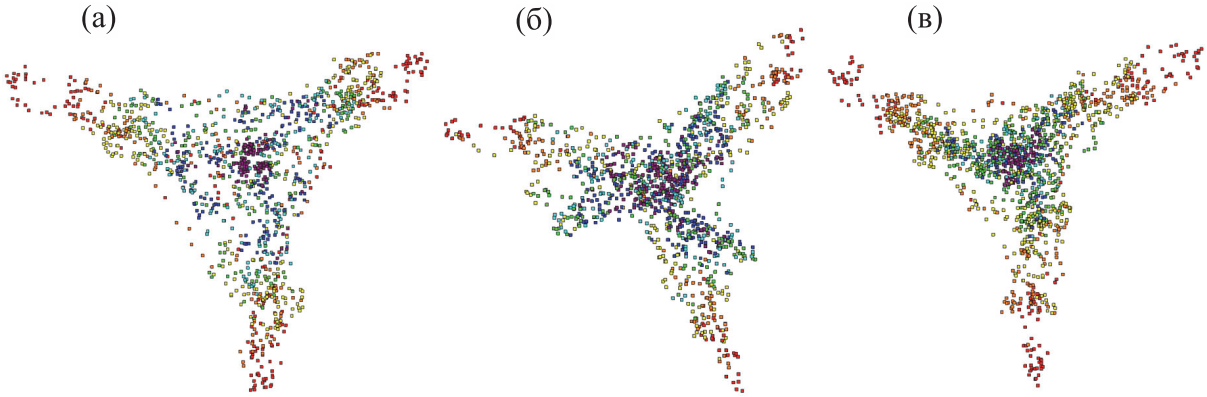

Рис. 6. Пространственное расположение GC-состава фрагментов генома *Homalodisca vitripennis* (рис. 6а), *Liphistius erawan* (рис. 6б) и *Trigoniophthalmus alternatus* (рис. 6в) в пространстве 1-й и 2-й главных компонент.

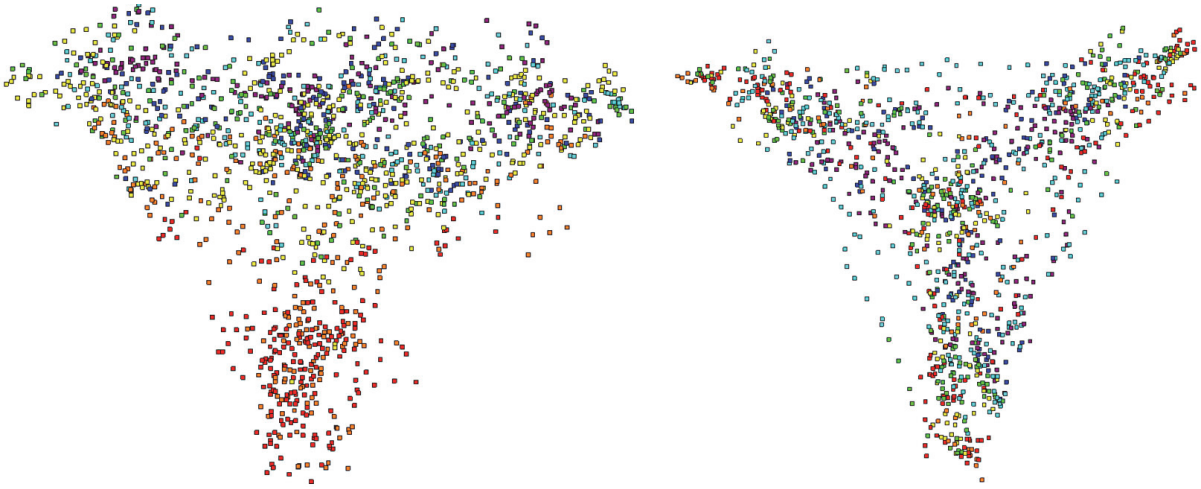

Рис. 7. Пространственное расположение GC-состава фрагментов генома *Callyspongia plicifera* (рис. 7а) в пространстве 1-й и 3-й главных компонент и *Echinococcus equinus* (рис. 7б) в пространстве 1-й и 2-й главных компонент.

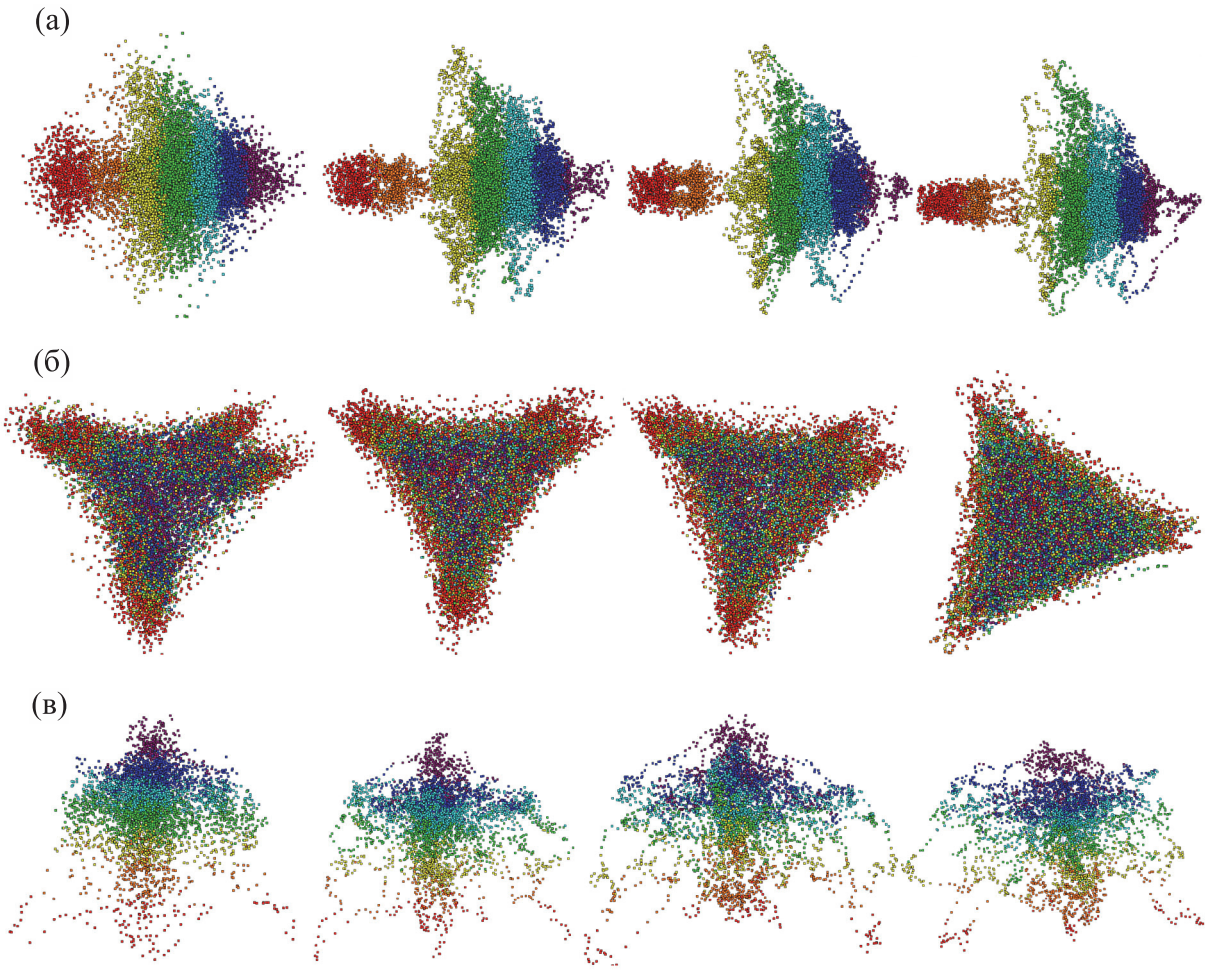

Рис. 8. Распределение величины GC-состава фрагментов геномов хлоропластов (а), бактерий (б) и митохондрий (в) для различной длины окна.

Таблица 1. Скалярные произведения векторов GC-состава триплетов на первые четыре главных компоненты

	1-я главная компонента	2-я главная компонента	3-я главная компонента	4-я главная компонента
Anthoceros angustus	**–0.10627**	0.003868	–0.0083	0.002669
Synechococcus sp. CC9311	**–0.06197**	–0.00552	0.000913	0.002057
Clostridium baratii str. Sullivan	–0.00253	–0.00733	–0.00808	**–0.11122**
Trigoniophthalmus alternatus	0.000549	–0.0011	–0.0455	**–0.06746**

ЧЕМ ОБУСЛОВЛЕНО ГРАДИЕНТНОЕ ИЛИ ЦЕНТРАЛЬНОСИММЕТРИЧНОЕ РАСПРЕДЕЛЕНИЕ ЗНАЧЕНИЙ GC-СОСТАВА

Рассмотрим скалярное произведение вектора, компонентами которого будут значения GC-состава всех триплетов, на вектора главных компонент. Построим 64-мерный вектор GC-состава триплетов следующим образом. Триплету AAA будет соответствовать значение GC-состава 0, триплету AAC будет соответствовать значение GC-состава 1, триплету ACC – значение 2 и т.д. Нормируем этот вектор, чтобы сумма всех компонент была равна 1. Анализ скалярного произведения показал, что при градиентном распределении значений GC-состава максимальное по абсолютной величине значение скалярного произведения наблюдается при умножении вектора GC-состава триплетов на первую или вторую главную компоненты. Для центральносимметричного распределения максимальное по абсолютной величине значение скалярного произведения наблюдается при умножении вектора GC-состава триплетов на третью или четвертую главные компоненты. В табл. 1 в качестве примера приведены скалярные произведения векторов GC-состава триплетов на первые четыре главных компоненты. Жирным шрифтом выделены максимальные по абсолютной величине значения. *Anthoceros angustus* и *Synechococcus* sp. *CC9311* относятся к градиентному распределению значений GC-состава, для них максимальное значение скалярного произведения получено при умножении на первую главную компоненту. *Clostridium baratii str. Sullivan* и *Trigoniophthalmus alternatus* относятся к центральносимметричному распределению, для них максимальное значение скалярного произведения получено при умножении на четвертую главную компоненту.

ВЫВОДЫ

Показано существование упорядоченности в распределении значений GC-состава фрагментов различных геномов. При этом данная упорядоченность имеет типичный вид для отдельных групп геномов. Наибольшую однородность распределения величины GC-состава показали геномы хлоропластов наземных растений. Для них характерно градиентное распределение. Тип распределения для геномов бактерий обусловлен величиной GC-состава генома в целом. Для геномов, величина GC-состава которых менее 50%, наблюдается центральносимметричное распределение. Для геномов с величиной GC-состава больше 50% распределение носит градиентный характер. Наибольшим разнообразием в типах распределений GC-состава фрагментов обладают геномы митохондрий. Тем не менее, можно проследить некоторые закономерности и для этих геномов. Так, для геномов митохондрий растений (наземных растений, одноклеточных и многоклеточных водорослей, обычных и печеночных мхов), а также для высших и низших грибов характерно градиентное распределение. Для геномов митохондрий насекомых, паукообразных и ракообразных характерно центральносимметричное распределение. Для всех остальных групп геномов митохондрий животных наблюдались оба типа распределений в каждой группе. Возможно, это связано с более сложным строением организмов и разнообразием среды их обитания.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая статья не содержит описания каких-либо исследований с участием людей или животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. Shimda H. and Sugiuro M. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. *Nucl. Acids Res.*, **19** (5), 983–995 (1991). DOI: 10.1093/nar/19.5.983

2. Young H. A., Lanzatella C. L., Sarath G., and Tobias C. M. Chloroplast genome variation in upland and lowland switchgrass. *PLoS One*, **6** (8), e23980 (2011). DOI: 10.1371/journal.pone.0023980
3. Lockhart P. J., Penny D., Hendy M. D., Howe C. J., Beanland T. J., and Larkum A. W. D. Controversy on chloroplast origins. *FEBS Lett.*, **301** (2), 127–131 (1992). DOI: 10.1016/0014-5793(92)81231-A
4. Gao L., Yi X., Yang Y. X., Su Y. J., and Wang T. Complete chloroplast genome sequence of a tree fern *Alsophila spinulosa*: insights into evolutionary changes in fern chloroplast genomes. *BMC Evol. Biol.*, **9**, 130 (2009). DOI: 10.1186/1471-2148-9-130
5. Wu Z. Q. and Ge S. The phylogeny of the BEP clade in grasses revisited: evidence from the whole-genome sequences of chloroplasts. *Mol. Phylogen. Evol.*, **62** (1), 573–578 (2012). DOI: 10.1016/j.ympev.2011.10.019
6. Qian J., Song J., Gao H., Zhu Y., Xu J., Pang X., Yao H., Sun C., Li X., Li C., Liu J., Xu H., and Chen S. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PLoS One*, **8** (2), e57607 (2013). DOI: 10.1371/journal.pone.0057607
7. Zhang T., Fang Y., Wang X., Deng X., Zhang X., Hu S., and Yu J. The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. *PLoS One*, **7** (1), e30531 (2012). DOI: 10.1371/journal.pone.0030531
8. Yang Y., Zhou T., Duan D., Yang J., Feng L., and Zhao G. Comparative analysis of the complete chloroplast genomes of five Quercus species. *Front. Plant Sci.*, **7**, 959 (2016). DOI: 10.3389/fpls.2016.00959
9. Hildebrand F., Meyer A., and Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genetics*, **6** (9), e1001107 (2010). DOI: 10.1371/journal.pgen.1001107
10. Gorban A. N., Popova T. G., and Zinovyev A. Y. Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences. *In silico Biol.*, **5** (3), 265–282 (2005). DOI: 10.48550/arXiv.q-bio/0410033
11. Gorban A. N., Zinovyev A. Y., and Popova T. G. Seven clusters in genomic triplet distributions. *In silico Biol.*, **3** (4), 471–482 (2003). DOI: 10.48550/arXiv.cond-mat/0305681
12. Lightfield J., Fram N. R., and Ely B. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One*, **6** (3), e17677 (2011). DOI: 10.1371/journal.pone.0017677
13. Peano C., Pietrelli A., Consolandi C., Rossi E., Petiti L., Tagliabue L., De Bellis D., and Landini P. An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb. Informatics Exp.*, **3**, 1 (2013). DOI: 10.1186/2042-5783-3-1
14. Zhou H. Q., Ning L. W., Zhang H. X., and Guo F. B. Analysis of the relationship between genomic GC content and patterns of base usage, codon usage and amino acid usage in prokaryotes: similar GC content adopts similar compositional frequencies regardless of the phylogenetic lineages. *PLoS One*, **9** (9), e107319 (2014). DOI: 10.1371/journal.pone.0107319
15. Giannoukos G., Ciulla D. M., Huang K., Haas B. J., Izard J., Levin J. Z., Livny J., Earl A. M., Gevers D., Ward D. V., Nusbaum C., Birren B. W., and Gnirke A. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.*, **13**, r23 (2012). DOI: 10.1186/gb-2012-13-3-r23
16. Behura S. K., Lobo N. F., Haas B., DeBruyn B., Lovin D. D., Shumway M. F., Puiu D., Romero-Severson J., Nene V., and Severson D. W. Complete sequences of mitochondria genomes of Aedes aegypti and Culex quinquefasciatus and comparative analysis of mitochondrial DNA fragments inserted in the nuclear genomes. *Insect Biochem. Mol. Biol.*, **41** (10), 770–777 (2011). DOI: 10.1016/j.ibmb.2011.05.006
17. Johnston I. G. and Williams B. P. Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. *Cell Systems*, **2** (2), 101–111 (2016). DOI: 10.1101/037960
18. Ferla M. P. et al. New rRNA gene-based phylogenies of the Alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability. *PLoS One*, **8** (12), e83383 (2013). DOI: 10.1371/journal.pone.0083383
19. Nakamura Y., Sasaki N., Kobayashi M., Ojima N., Yasuike M., Shigenobu Y., Satomi M., Fukuma Y., Shiwaku K., Tsujimoto A., Kobayashi T., Nakayama I., Ito F., Nakajima K., Sano M., Wada T., Kuhara S., Inouye K., Gojobori T., and Ikeo K. The first symbiont-free genome sequence of marine red alga, Susabi-nori (*Pyropia yezoensis*). *PLoS One*, **8** (3), e57122 (2013). DOI: 10.1371/journal.pone.0057122
20. Godel C., Kumar S., Koutsovoulos G., Ludin P., Nilsson D., Comandatore F., Wrobel N., Thompson M., Schmid C.D., Goto S., BringaudF., Wolstenholme A., Bandi C., Epe C., Kaminsky R., Blaxter M., and Mäser P. The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets. *FASEB J.*, **26** (11), 4650 (2012). DOI: 10.1096/fj.12-205096
21. Imanian B., Pombert J. F., Dorrell R. G., Burki F., and Keeling P. J. Tertiary endosymbiosis in two dinotoms has generated little change in the mitochondrial genomes of their dinoflagellate hosts and diatom endosymbionts. *PLoS One*, **7** (8), e43763 (2012). DOI: 10.1371/journal.pone.0043763
22. Wei L., He J., Jia X., Qi Q., Liang Z., Zheng H., Ping Y., Liu S, and Sun J. Analysis of codon usage bias of mitochondrial genome in Bombyx mori and its relation to evolution *BMC Evol. Biol.*, **14**, 262 (2014). DOI: 10.1186/s12862-014-0262-4
23. Sadovsky M. G., Senashova M. Y., and Malyshev A. V. Amazing symmetrical clustering in chloroplast genomes. *BMC Bioinformatics*, **21** (Suppl. 2), 83 (2020). DOI: 10.1186/s12859-020-3350-z

Distribution of Values of GC-Content of the Fragments in the Spatial Structure of Mitochondrial, Chloroplast and Bacterial Genomes

M.Yu. Senashova* and M.G. Sadovsky*, **, ***

**Institute of Computational Modelling, Siberian Branch of the Russian Academy of Sciences, Akademgorodok 50/44, Krasnoyarsk, 660036*

***Federal Siberian Research Clinical Center, FMBA of Russia, Kolomenskaya ul. 26, Krasnoyarsk, 660037 Russia*

****Siberian Federal University, Svobodny prosp. 79, Krasnoyarsk, 660041 Russia*

Distribution of values of GC-content of the fragments in spatial structure of chloroplast, mitochondrial, and bacterial genomes were explored. It was found that GC-content in the fragments for most genomes is identically but not independent distributed varaible. Two main types of distribution have been revealed: the gradient distribution and centrally symmetrical distribution. Chloroplast genomes have only a gradient distribution. In bacteria, for the GC-poor genomes, a centrally symmetrical distribution is observed, while there is a gradient distribution in the GC-rich genomes. In mitochondria, both types of distribution are present, the type of distribution depends on the type of an organism.

Keywords: GC-content, frequency dictionaries, triplets, structure, spatial distribution