УДК 57.015 + 573.2
DOI:10.25729/ESI.2024.34.2.003

Упорядоченность значений GC-состава фрагментов в пространственной структуре геномов органелл

Сенашова Мария Юрьевна

Институт вычислительного моделирования СО РАН,
Россия, Красноярск, *msen@icm.krasn.ru*

Аннотация. Рассмотрено пространственное распределение значений GC-состава фрагментов геномов хлоропластов и митохондрий. Под пространственным распределением понимается распределение точек, соответствующих участкам геномов, в пространстве частот триплетов. Обнаружено, что значения GC-состава фрагментов для большинства геномов распределены не хаотически, а упорядоченно. Были обнаружены 2 основных типа распределения: градиентное и центрально-симметричное. У геномов хлоропластов встречается только градиентное распределение. У митохондрий встречаются оба типа распределения. Тип распределения для митохондрий зависит от вида организма. Пространственное распределение GC-состава является устойчивым относительно изменения длины окна считывания.

Ключевые слова: порядок, пространственное распределение, триплеты, частотные словари, главные компоненты

Цитирование: Сенашова М.Ю. Упорядоченность значений GC-состава фрагментов в пространственной структуре геномов органелл / М.Ю. Сенашова // Информационные и математические технологии в науке и управлении. – 2024. – № 2(34). – С. 33-40. – DOI:10.25729/ESI.2024.34.2.003.

Введение. Изучение особенностей и деталей структуры нуклеотидных последовательностей является важнейшей задачей биологии в настоящее время. Нуклеотидные последовательности являются очень интересным объектом исследования как с точки зрения биологии, так и с точки зрения биоинформатики. Биологи и биоинформатики заинтересованы в обнаружении упорядоченных структур в этих последовательностях. Однако, биология изучает нуклеотидные последовательности с точки зрения их функционирования, выявления тех или иных физических структур, их физико-химического взаимодействия, а биоинформатика рассматривает нуклеотидные последовательности как символьные и анализирует их с использованием математических методов.

Такой параметр, как GC-состав очень часто используется в исследованиях структуры геномов и функций отдельных участков. Этому посвящено много работ, как в случае геномов хлоропластов [1-8], так и геномов митохондрий [9-15]. Под GC-составом понимают долю гуанина и цитозина по отношению к длине рассматриваемой нуклеотидной последовательности. Пара гуанин-цитозин соединена тремя водородными связями, тогда как пара аденин–тимин — двумя. Это обуславливает разные физические свойства у GC-бедных и GC-богатых нуклеотидных последовательностей.

В данной работе нуклеотидные последовательности рассматриваются с точки зрения биоинформатики. Изучается распределение значений GC-состава выделенных фрагментов геномов хлоропластов и митохондрий 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.

1. Материалы и методы. Введем основные понятия. Мы будем рассматривать генетическую последовательность длины L, состоящую из символов алфавита $\aleph=\{A,C,G,T\}$. Для этой последовательности мы будем составлять частотный словарь толщины 3. Частотный

словарь W_3 толщины 3 символьной последовательности – это список всех троек $\omega = \nu_1\nu_2\nu_3$ идущих подряд нуклеотидов с указанием частот этих троек; всего может быть 64 триплета. Мы используем частотный словарь, в котором триплеты подсчитываются таким образом, что они полностью покрывают последовательность и при этом не пересекаются. Частота f_ω – это отношение числа копий n_ω данного слова к общему числу всех триплетов N, где N – сумма всех n_ω:

$$f_\omega = \frac{n_\omega}{N} \tag{1}$$

Всякий частотный словарь W_3 отображает геном в 64-мерное метрическое пространство.

Один из 64 триплетов исключался, поскольку сумма всех частот в словаре равна 1, что порождает линейную связь, которая будет давать ложный сигнал при статистической обработке (корреляционном анализе, определении главных компонент и т.п.).

В нашем случае целесообразнее исключать тот триплет, для которого стандартное отклонение, наблюдаемое по анализируемому набору 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.

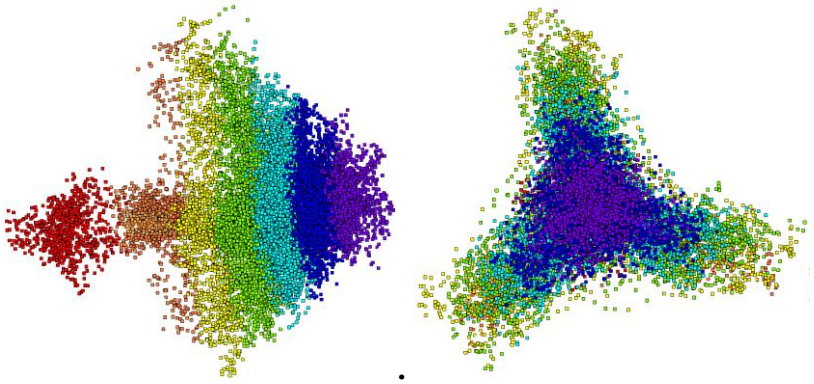
Для выявления структуры в генетической последовательности проводилась предварительная обработка, которая ставила в соответствие данной последовательности множество точек в 63-мерном пространстве триплетов. Делалось это следующим образом: последовательность сканировалась рамкой считывания длины Δ с шагом t. Для каждого положения i рамки определялся участок генетической последовательности, совпадающий с рамкой считывания, для которого вычислялся частотный словарь $W_3^{(i)}$ соответствующий i-ой точке в 63-мерном пространстве. Кроме того, с каждой точкой в 63-мерном пространстве связывались следующие параметры: номер центрального символа рассматриваемого участка и величина GC-состава этого участка.

Данные для исследования брались в базе EMBL-банка. Было отобрано 418 геномов митохондрий растений и животных, в среднем по 25 видов на кладу и 391 геном хлоропластов наземных растений. Для всех генетических последовательностей длина рамки считывания $\Delta = 603$, шаг $t = 11$.

По полученному множеству точек в программе VidaExpert (http://bioinfo-out.curie.fr/projects/vidaexpert/) строился вид данных в пространстве первых трех главных компонент, вычисленных для данного 63-мерного пространства. Рассматривались две проекции на плоскость пространства главных компонент: 1-ой и 2-ой компоненты и 1-ой и 3-ей компоненты или 2-ой и 3-ей компоненты, в зависимости от структуры генома. В обеих проекциях точки окрашивались в соответствии со значениями GC-состава. Окраска выполнялась следующим образом. Строилась гистограмма по величине GC-состава. Все точки в интервале изменения GC-состава делились на семь примерно равных частей пропорционально количеству точек с соответствующими значениями. Точки, распределенные по семи интервалам, раскрашивались в соответствии с цветами радуги. Минимальным значениям GC-состава соответствует фиолетовый цвет, максимальным – красный. Указанные цвета используются для всех рисунков ниже по тексту.

2. Пространственная структура величины GC-состава фрагментов геномов хлоропластов. Подавляющее большинство геномов хлоропластов наземных растений имеет

трехлучевую структурут генома [16]. Кроме того, в пространстве первой и второй главных компонент виден отдельный кластер, так называемый «хвост» (рис. 1а слева). На рис 1 а, б показана структура генома хлоропласта с раскрашенным распределением величины *GC* - состава на примере генома *Anthoceros angustus* (идентификатор AB086179 в EMBL). Исключен триплет *gcg*. Было обнаружено, что распределение величины *GC* -состава фрагментов генома 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. 1, а).



а) б)

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

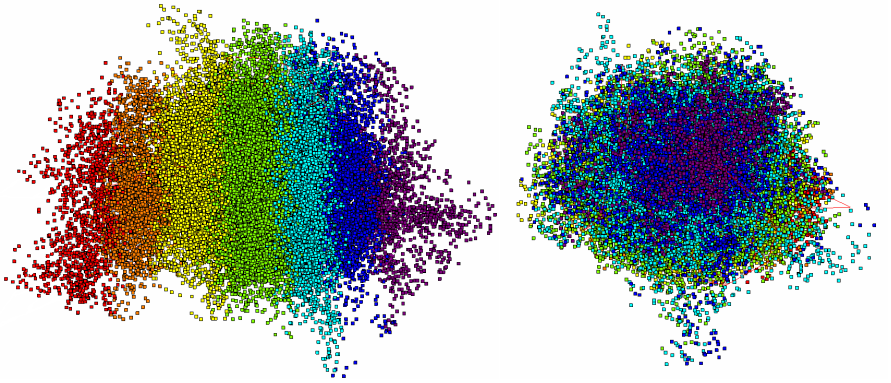
3. Пространственная структура величины *GC* -состава фрагментов геномов митохондрий. Пространственная структура геномов митохондрий имеет преимущественно трехлучевую структуру кроме наземных растений и печеночных мхов, которые имеют структуру близкую к эллипсоиду.

Рассмотрим группы геномов, для которых наблюдается однотипное распределение значений *GC* -состава по пространственной структуре геномов. Рассмотрим сначала геномы митохондрий 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. 2 показано распределение значений *GC* -состава *Zea mays* (идентификатор AB251495 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 всех остальных геномов митохондрий характерна трехлучевая структура и далее отдельно это указываться не будет.

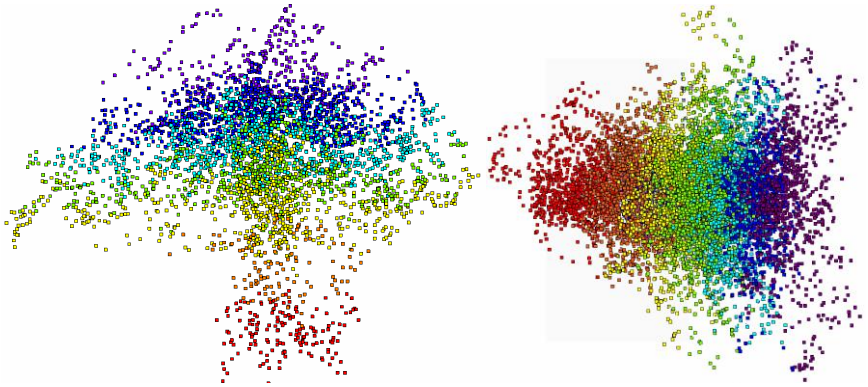
Для одноклеточных водорослей, обычных мхов, мхоподобных лишайников из рода *Cladonia* и высших грибов характерно градиентное распределение значений *GC* -состава фрагментов вдоль оси симметрии структуры генома. На рис. 3 показано распределение значений *GC* -состава фрагментов на примере *Mesostigma viride* (идентификатор AF353999 в EMBL, исключен триплет *ccg*), *Physcomitrella patens* (идентификатор AY506529 в EMBL, исключен триплет *ccg*), *Cladonia petrophila* (идентификатор MG941021 в EMBL, исключен

триплет *ccg*) и *Agaricus bisporus var. bisporus H97* (идентификатор MG941021 в EMBL, исключен триплет *gcg*).

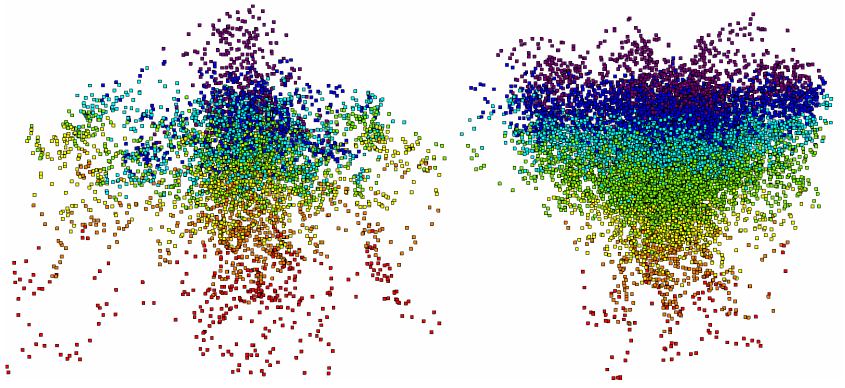


а) б)

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



а) б)

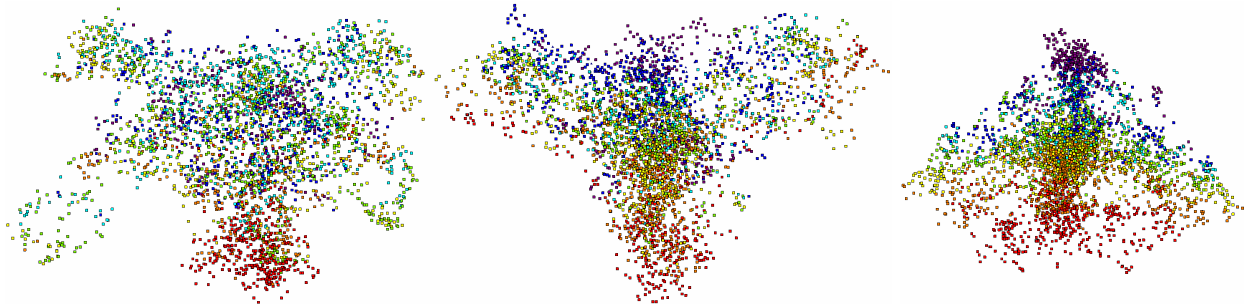


в) г)

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

Многоклеточные водоросли, лишайники и низшие грибы не имеют ярко выраженного градиентного распределения. 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. 4 на примере *Saccharina japonica* (идентификатор AP011493 в EMBL,

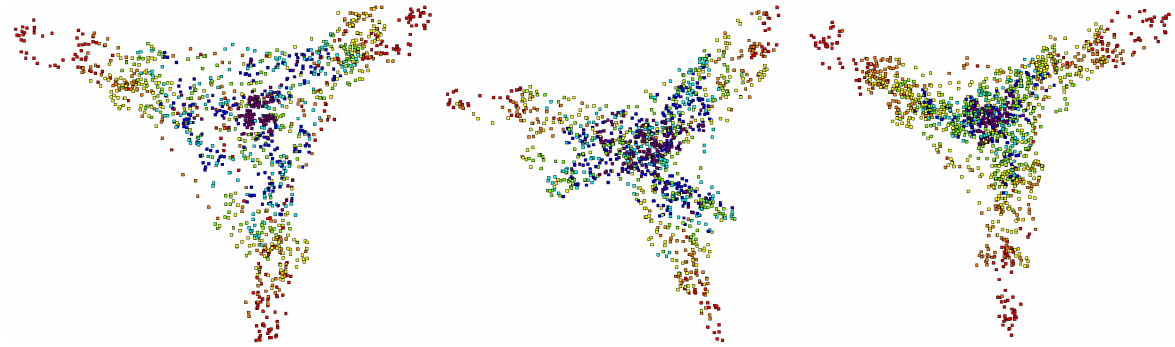
исключен триплет *cgc*), *Hypogymnia vittata* (идентификатор KY362374 в EMBL, исключен триплет *cgc*), *Candida viswanathii* (идентификатор EF536359 в EMBL, исключен триплет *ccg*).



а) б) в)

Рис. 4. Пространственное расположение *GC* -состава фрагментов генома *Saccharina japonica* (рис. 4а), *Hypogymnia vittata* (рис. 4б) и *Candida viswanathii* (рис. 4в) в пространстве 1 и 3 главных компонент

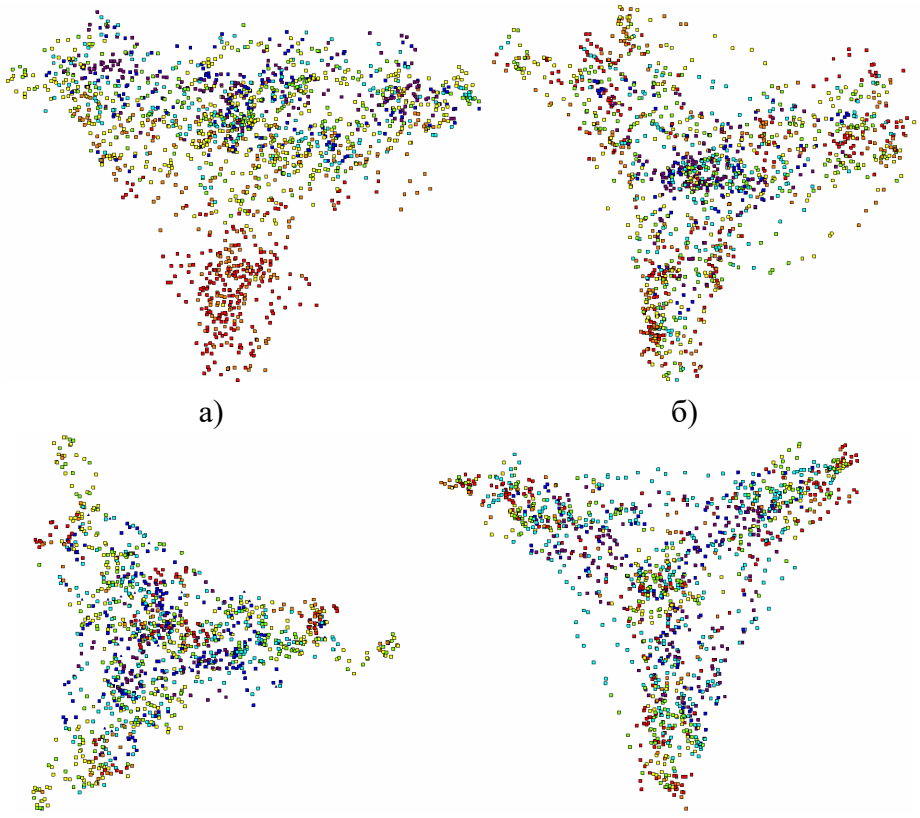
Перейдем к рассмотрению геномов митохондрий животных. Для геномов митохондрий насекомых, паукообразных и ракообразных характерно центральносимметричное распределение значений *GC* -состава фрагментов по пространственной структуре геномов. Минимальные значения *GC* -состава расположены в центре трехлучевой 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. 5 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*).



а) б) в)

Рис. 5. Пространственное расположение *GC* -состава фрагментов генома *Homalodisca vitripennis* (рис. 5а), *Liphistius erawan* (рис. 5б) и *Trigoniophthalmus alternatus* (рис. 5в) в пространстве 1 и 2 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. 6 представлены все варианты на примере *Callyspongia plicifera* (идентификатор EU237477 в EMBL, исключен триплет *cgc*), *Neritina usnea* (идентификатор KU342665 в EMBL, исключен триплет *cgc*) *Urechis caupo* (идентификатор AY619711 в EMBL, исключен триплет *gcg*), *Echinococcus equinus* (идентификатор AF346403 в EMBL, исключен триплет *cgc*).



а) б)

в) г)

Рис. 6. Пространственное расположение *GC* -состава фрагментов генома *Callyspongia plicifera* (рис. 6а) в пространстве 1 и 3 главных компонент., *Neritina usnea* (рис. 6б), *Urechis caupo* (рис. 6в) и *Echinococcus equinus* (рис. 6г) в пространстве 1 и 2 главных компонент

Для позвоночных животных не удалось выявить какую-либо типичную картину распределения *GC* -состава фрагментов внутри классов и между классами.

Заключение. Результаты, описанные в статье, показывают, что существует упорядоченность в распределении значений *GC* -состава фрагментов различных геномов. Причем эта упорядоченность имеет типичный вид для отдельных групп геномов. Геномы митохондрий обладают большим разнообразием типов распределений *GC* -состава фрагментов. Это обусловлено тем, что митохондрии, рассмотренные в этой работе, принадлежат различным классам организмов с отличающимся строением и функционалом. В то время как геномы хлоропластов достаточно однородны. Обнаруженная упорядоченность – это очень интересный результат, поскольку изначально *GC* -состав описывал физические свойства геномов, но как оказалось, он влияет и на статистические свойства геномов, рассматриваемых как символьные последовательности.

Список источников

1. Shimda H., Sugiuro M. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic acids research, 1991, vol. 19, no. 5, pp. 983-995.
2. Young H.A. et al. Chloroplast genome variation in upland and lowland switchgrass. PloS one, 2011, vol. 6, no. 8, pp. e23980.
3. Lockhart P.J., Penny D., Hendy M.D. et al. Controversy on chloroplast origins. FEBS letters, 1992, vol. 301, no. 2, pp. 127-131.
4. Gao L., Yi X., Yang Y.-X. et al. Complete chloroplast genome sequence of a tree fern Alsophila spinulosa: insights into evolutionary changes in fern chloroplast genomes. BMC evolutionary biology, 2009, vol. 9, no. 1, 130 p.
5. Wu Z.Q., Ge S. The phylogeny of the BEP clade in grasses revisited: evidence from the whole-genome sequences of chloroplasts. Molecular phylogenetics and evolution, 2012, vol. 62, no. 1, pp. 573-578.
6. Qian J., Song. J., Gao H. et al. The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza. PloS one, 2013, vol. 8, no. 2, pp. e57607.

7. Zhang T., Fang Y., Wang X. et al. The complete chloroplast and mitochondrial genome sequences of Boea hygrometrica: insights into the evolution of plant organellar genomes. PLoS One, 2012, vol. 7, no. 1, pp. e30531.
8. Yang Y., Zhou T., Duan D. et al. Comparative analysis of the complete chloroplast genomes of five Quercus species. Frontiers in plant science, 2016, vol. 7, 959 p.
9. Behura S.K., Lobo N.F., Haas Br. et al. Complete sequences of mitochondria genomes of Aedes aegypti and Culex quinquefasciatus and comparative analysis of mitochondrial DNA fragments inserted in the nuclear genomes. Insect biochemistry and molecular biology, 2011, vol. 41, no. 10, pp. 770-777.
10. Johnston I.G., Williams B. P. Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. Cell systems, 2016, vol. 2, no. 2, pp. 101-111.
11. Ferla M.P., Thrash J.C., Giovannoni St.J. et al. New rRNA gene-based phylogenies of the Alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability. PLoS One, 2013, vol. 8, no. 12, pp. e83383.
12. Nakamura Y., Sasaki N.V., Kobayashi M. et al. The first symbiont-free genome sequence of marine red alga, Susabi-nori (Pyropia yezoensis). PloS one, 2013, vol. 8, no. 3, pp. e57122.
13. Godel C. et al. The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets. The FASEB Journal, 2012, vol. 26, no. 11, pp. 4650-4661.
14. Imanian B., Pombert J.-F., Dorrell R. et al. Tertiary endosymbiosis in two dinotoms has generated little change in the mitochondrial genomes of their dinoflagellate hosts and diatom endosymbionts. PLoS One, 2012, vol. 7, no. 8, pp. e43763.
15. Wei L. et al. Analysis of codon usage bias of mitochondrial genome in Bombyx mori and its relation to evolution. BMC evolutionary biology, 2014, vol. 14, no. 1, 262 p.
16. Sadovsky M.G., Senashova M.Yu., Malyshev A.V. Amazing symmetrical clustering in chloroplast genomes. BMC Bioinformatics, 2020, 21(Suppl 2):83.

Сенашова Мария Юрьевна. *ИВМ СО РАН, к.ф.-м.н., старший научный сотрудник. Направления исследований: анализ данных, биоинформатика. AuthorID: 133140, SPIN: 1178-6320, ORCID: 0000–0002–1023–7103, msen@icm.krasn.ru, 660036, Красноярск, Академгородок 50, стр.44.*

UDC 57.015 + 573.2
DOI:10.25729/ESI.2024.34.2.003

Orderliness of GC-content values of fragments in the spatial structure of organelle genomes

Maria Yu. Senashova

Institute of computational modelling of the SB RAS,
Russia, Krasnoyarsk, *msen@icm.krasn.ru*

Abstract. The spatial distribution of GC-content values of chloroplast and mitochondrial genome fragments is considered. Spatial distribution refers to the distribution of points corresponding to genome regions in the frequency space of triplets. It was found that the values of the GC-content of fragments for most genomes are distributed not chaotically, but in an orderly manner. Two main types of distribution were found: gradient and centrally symmetric. In chloroplast genomes, only a gradient distribution occurs. In mitochondria genomes, both types of distribution occur. The type of distribution for mitochondria genomes depends on the type of organism. The spatial distribution of the GC-content is stable with respect to changes in the reading window length.

Keywords: order, spatial distribution, triplets, frequency dictionaries, principal components

References

1. Shimda H., Sugiuro M. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic acids research, 1991, vol. 19, no. 5, pp. 983-995.
2. Young H.A. et al. Chloroplast genome variation in upland and lowland switchgrass. PloS one, 2011, vol. 6, no. 8, pp. e23980.

3. Lockhart P.J., Penny D., Hendy M.D. et al. Controversy on chloroplast origins. FEBS letters, 1992, vol. 301, no. 2, pp. 127-131.
4. Gao L., Yi X., Yang Y.-X. et al. Complete chloroplast genome sequence of a tree fern Alsophila spinulosa: insights into evolutionary changes in fern chloroplast genomes. BMC evolutionary biology, 2009, vol. 9, no. 1, 130 p.
5. Wu Z.Q., Ge S. The phylogeny of the BEP clade in grasses revisited: evidence from the whole-genome sequences of chloroplasts. Molecular phylogenetics and evolution, 2012, vol. 62, no. 1, pp. 573-578.
6. Qian J., Song. J., Gao H. et al. The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza. PloS one, 2013, vol. 8, no. 2, pp. e57607.
7. Zhang T., Fang Y., Wang X. et al. The complete chloroplast and mitochondrial genome sequences of Boea hygrometrica: insights into the evolution of plant organellar genomes. PLoS One, 2012, vol. 7, no. 1, pp. e30531.
8. Yang Y., Zhou T., Duan D. et al. Comparative analysis of the complete chloroplast genomes of five Quercus species. Frontiers in plant science, 2016, vol. 7, 959 p.
9. Behura S.K., Lobo N.F., Haas Br. et al. Complete sequences of mitochondria genomes of Aedes aegypti and Culex quinquefasciatus and comparative analysis of mitochondrial DNA fragments inserted in the nuclear genomes. Insect biochemistry and molecular biology, 2011, vol. 41, no. 10, pp. 770-777.
10. Johnston I.G., Williams B. P. Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. Cell systems, 2016, vol. 2, no. 2, pp. 101-111.
11. Ferla M.P., Thrash J.C., Giovannoni St.J. et al. New rRNA gene-based phylogenies of the Alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability. PLoS One, 2013, vol. 8, no. 12, pp. e83383.
12. Nakamura Y., Sasaki N.V., Kobayashi M. et al. The first symbiont-free genome sequence of marine red alga, Susabi-nori (Pyropia yezoensis). PloS one, 2013, vol. 8, no. 3, pp. e57122.
13. Godel C. et al. The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets. The FASEB Journal, 2012, vol. 26, no. 11, pp. 4650-4661.
14. Imanian B., Pombert J.-F., Dorrell R. et al. Tertiary endosymbiosis in two dinotoms has generated little change in the mitochondrial genomes of their dinoflagellate hosts and diatom endosymbionts. PLoS One, 2012, vol. 7, no. 8, pp. e43763.
15. Wei L. et al. Analysis of codon usage bias of mitochondrial genome in Bombyx mori and its relation to evolution. BMC evolutionary biology, 2014, vol. 14, no. 1, 262 p.
16. Sadovsky M.G., Senashova M.Yu., Malyshev A.V. Amazing symmetrical clustering in chloroplast genomes. BMC Bioinformatics, 2020, 21(Suppl 2):83.

Senashova Maria Yurievna. *ICM SB RAS, Ph.D., senior researcher. Research direction: data analysis, bioinformatics. AuthorID: 133140, SPIN: 1178-6320, ORCID: 0000–0002–1023–7103, msen@icm.krasn.ru, 660036, Krasnoyarsk, Akademgorodok 50, building 44.*

Статья поступила в редакцию 08.12.2023; одобрена после рецензирования 30.05.2024; принята к публикации 03.06.2024.
The article was submitted 12/08/2023; approved after reviewing 05/30/2024; accepted for publication 06/03/2024.