УДК 573.22:575.89

СТРУКТУРА ГЕНОМОВ ХЛОРОПЛАСТОВ ВОДОРОСЛЕЙ

[1]Сенашова М.Ю., [1,2]Садовский М.Г.

[1]ФГБУН «Федеральный исследовательский центр «Красноярский научный центр Сибирского отделения Российской академии наук»» – обособленное подразделение «Институт вычислительного моделирования Сибирского отделения Российской академии наук», Красноярск, e-mail: msen@icm.krasn.ru;

[2]ФГАОУ «Сибирский федеральный университет», Институт фундаментальной биологии и биотехнологии, Красноярск, e-mail: msad@icm.krasn.ru

В работе изложены результаты статистического анализа структуры геномов хлоропластов водорослей. В нашей работе в качестве структуры рассматривается расположение точек, соответствующих участкам генома, в 63-мерном пространстве частот троек символов. Геном сканируется окном длины Δ со сдвигом t. Для каждого полученного таким образом участка длины Δ вычисляется частотный словарь троек символов. Тройки символов в каждом участке рассматривались без пересечения. Было проанализировано 11 геномов хлоропластов многоклеточных водорослей и 65 геномов хлоропластов одноклеточных водорослей. Визуализация структуры каждого генома была сделана в программе VidaExpert. Для этого была построена проекция 64-мерного пространства частот в пространство первых трёх главных компонент. Исследования показали, что большинство геномов хлоропластов водорослей в пространстве первых трёх главных компонент имеют характерную структуру. Однако для геномов хлоропластов водорослей наблюдается существенно больше вариантов, отличающихся от характерной структуры, чем для геномов хлоропластов наземных растений.

Ключевые слова: тройка символов, частота, структура данных, геном

STRUCTURE OF GENOMES OF CHLOROPLAST OF SOME ALGAE

[1]Senashova M.Yu., [1,2]Sadovskiy M.G.

[1]Institute of Computational Modeling of Siberian Branch of Russian Academy of sciences, Krasnoyarsk, e-mail: msen@icm.krasn.ru;

[2]Siberian Federal University, Institute of Fundamental Biology and Biotechnology, Krasnoyarsk, e-mail: msad@icm.krasn.ru

Some preliminary results on the chloroplast genomes structure of algae are present. Structure here is a pattern of the distribution of a set of points in 63-dimenstional metric space, where each point is the frequency dictionary of a fragment of a chromosome of the length Δ nucleotides identified with a step in t nucleotides. Each fragment has been converted into triplet frequency dictionary, where the triplets cover the fragments in non-overlapping manner, but with no gaps. 11 genomes of chloroplasts of multicellular algae and 65 genomes of unicellular algae were studied. With VidaExpert software the clustering of the frequency dictionaries corresponding to the fragments was obtained. In the space of three largest principal components the patterns exhibit significant likelihood.

Keywords: triplet, frequency, data pattern, taxonomy

Одной из важнейших задач в настоящее время для генетики и биоинформатики является определение структурных единиц в геномах как организмов в целом, так и отдельных органелл. Не менее важной задачей является выявление связи как между самими структурами, так и между структурами и выполняемыми ими функциями. В работах Горбаня с соавторами [1, 2] было показано, что для геномов бактерий характерна семикластерная структура. У бактерий участки генома группируются в соответствии с принадлежностью к кодирующим и некодирующим областям. Конфигурация кластеров в пространстве зависит от *GC*-состава генома, но их количество остается неизменным.

Изучение геномов органелл существенно помогает в получении ответа на вопрос о связи структуры генома и таксономии; в настоящее время общепринятой является теория, согласно которой хлоропласты растений произошли от бактерий. Большой интерес исследователей до сих пор вызывает сама теория [3–5], эволюция хлоропластов [6–8] и происхождение растений [10, 11]. Поэтому особенный интерес представляет изучение геномов тех бактерий, которые могут иметь общих с хлоропластами предков (в частности, цианобактерий), и определение подобия в структурах их геномов [12]. В рамках этой работы исследованы структуры геномов хлоропластов одноклеточных и многоклеточных водорослей. Выбор в качестве объектов изучения именно хлоропластов определяется в первую очередь тем, что они выполняют одну и ту же функцию. Кроме того размер геномов хлоропластов достаточно небольшой ($\approx 10^5$ символов). Основной задачей данной работы было определение особен-

ностей, свойственных именно геномам хлоропластов водорослей и их сравнение со структурами, полученными таким же методом для геномов других организмов.

Материалы и методы исследования

Введём понятия, используемые далее в работе. Мы будем рассматривать генетическую последовательность длины L, состоящую из символов алфавита $\mathfrak{M} = \{A, C, G, T\}$. Если последовательность содержит символы, отличающиеся от символов алфавита $\mathfrak{M}$, то такие символы из последовательности удаляются, а длина последовательности уменьшается на число таких символов. Под частотным словарем данной генетической последовательности будем понимать множество всех троек символов $\nu_1\nu_2\nu_3$ идущих подряд символов с соответствующими им частотами. Общее число таких троек равно 64. Отношение количества копий n_ω данной тройки символов к общему числу всех троек будем называть частотой:

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

Частотный словарь задает отображение генома в 64-мерное метрическое пространство, состоящее из точек, соответствующих частотным словарям участков генома. Два генома близки, если расстояние между множествами соответствующих им точек в Евклидовой метрике мало. Евклидова метрика для двух словарей задается следующим образом:

$$\rho\left(W_3^{(1)}, W_3^{(2)}\right) = \sqrt{\sum_{\omega=\mathrm{AAA}}^{\mathrm{TTT}} \left(f_\omega^{(1)} - f_\omega^{(2)}\right)^2}. \qquad (2)$$

Геном сканировался окном длины $\Delta = 603$ со сдвигом $t = 11$. Каждый участок длины Δ разбивался на тройки символов без пересечений, и для этого участка вычислялся частотный словарь. Таким образом, каждому участку генома ставилась в соответствие точка в 64-мерном пространстве, координатами которой являются частоты троек символов, входящие в участок. Для исключения влияния линейной зависимости между тройками символов (частоты в сумме дают единицу) одна из 64 троек символов удалялась из рассмотрения. Это снижает погрешность, которую линейная зависимость вносит в статистическую обработку данных. Выбор исключаемой тройки символов жестко не определен, но существуют эмпирические правила для выбора такой тройки. Например, можно исключать максимальную по значению частоты тройку символов, тем более если значение частоты этой тройки символов на порядок больше соответствующей величины следующей за ней тройки. Еще один подход предлагает в качестве исключаемой тройки символов выбирать тройку с минимальной величиной стандартного отклонения, вычисленного по множеству участков рассматриваемого генома. Тройка с таким стандартным отклонением оказывает наименьшее влияние на различимость объектов между собой (если стандартное отклонение равно 0, различия отсутствуют). В нашей работе мы пользовались вторым подходом. Минимальные значения стандартного отклонения наблюдались в основном для троек *GCG* и *CGC*, хотя встречались и другие тройки символов.

С каждой точкой в полученном после исключения одной из компонент 63-мерном пространстве связывались следующие параметры: номер центрального символа рассматриваемого участка и относительная фаза.

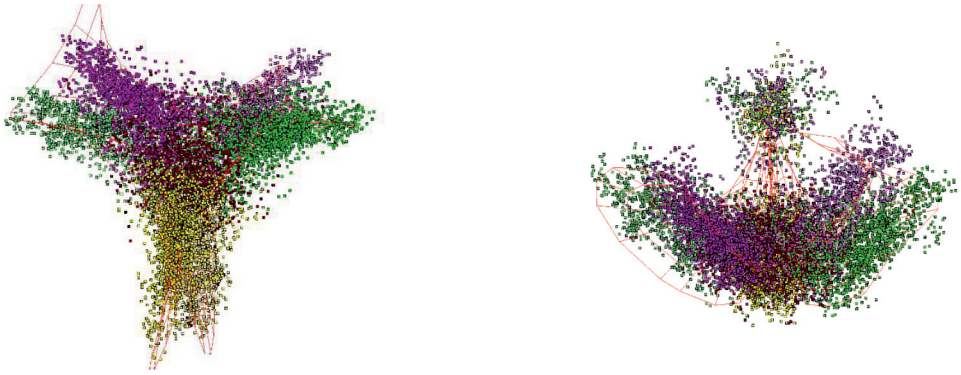
Номер центрального символа участка совпадает с номером этого символа в последовательности. Относительная фаза определяется с учетом того, попал рассматриваемый участок в кодирующую или некодирующую область последовательности. Участок относится к кодирующим, если он целиком попадал в кодирующую область последовательности. Если участок относится к некодирующим, то соответствующая ему точка помечается символом *J*. Для кодирующего участка возможны 6 вариантов маркировки: B_0, B_1, B_2, F_0, F_1, F_2. Если кодирующий участок в геноме аннотирован как считывающийся в прямом направлении, то для него вычислялся остаток от деления на 3 разности номеров центрального символа участка и первого символа кодирующей области, к которой он относится. В соответствии с величиной остатка от деления точка помечалась символами B_0, B_1 или B_2. Если участок аннотирован как считывающийся в обратном направлении, то вычислялся остаток от деления на 3 разности номеров последнего символа кодирующей области, к которой относится участок, и центрального символа участка. В зависимости от значения остатка от деления точка помечалась символами F_0, F_1 или F_2. Для всех генетических последовательностей длина рамки считывания $\Delta = 6003$, шаг $t = 101$.

Для того чтобы визуализировать множество точек 63-мерного пространства, с использованием программы *VidaExpert* [13] строилась проекция из 63-мерного пространства частот троек в пространство первых трёх главных компонент, построенных по этому множеству точек. Чтобы получить двумерные рисунки трехмерного пространства, рассматривались проекции на плоскость первых двух главных компонент и второй и третьей главной компоненты. Чтобы отобразить принадлежность точек к некодирующим областям и выделить относительные фазы, точки были помечены разными цветами. Точкам, относящимся к некодирующим областям, соответствует коричневый цвет. Для точек, относящихся к участкам с фазами B_0 и F_0, соответствуют темно-малиновый и светло-малиновый цвета, участкам с фазами B_1 и F_1 соответствуют темно-зеленый и светло-зеленый цвета, а к участкам с фазам B_2 и F_2 соответствуют темно-желтый и светло-желтый цвета. Для бактерий было показано, что *GC*-состав оказывает существенное влияние на расположение кластеров в пространстве первых трех главных компонент. Мы также вычисляли *GC*-состав геномов хлоропластов водорослей, чтобы выяснить, влияет ли этот параметр на структуру геномов хлоропластов.

Все исследованные геномы находятся в базе EMBL-банка.

Результаты исследования и их обсуждение

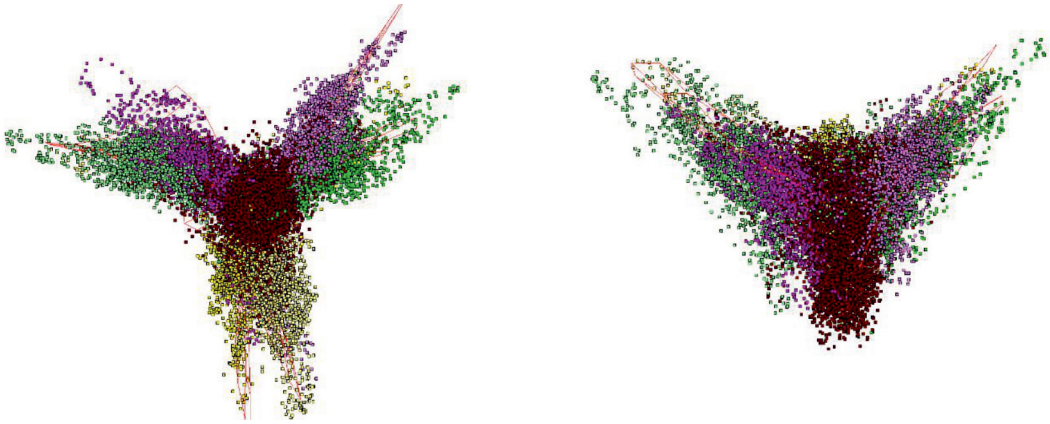
Исследование показало, что большая часть геномов хлоропластов водорослей (8 многоклеточных и 47 одноклеточных) имеет четкую трехлучевую структуру. Для этих геномов характерен вид в плоскости первой и второй главных компонент и плоскости второй и третьей главных компонент, который изображен на рис. 1.



а) Проекция в плоскость первых двух главных компонент

б) Проекция в плоскость второй и третьей главных компонент

Рис. 1. Характерный вид распределения участков хлоропластных геномов водорослей по частотам троек нуклеотидов в проекциях пространства первых трех главных компонент (приведена структура генома Phaeodactylum tricornutum)



а) Проекция в плоскость первых двух главных компонент

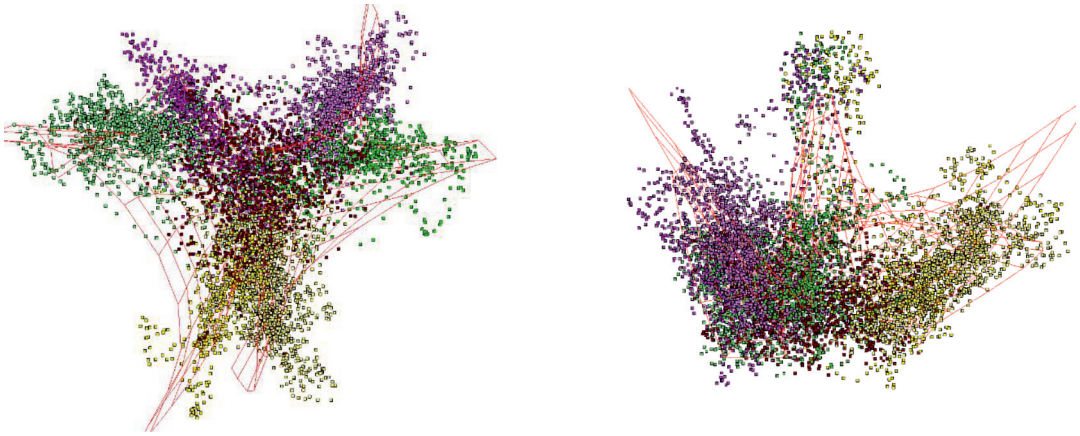
б) Проекция в плоскость второй и третьей главных компонент

Рис. 2. Структура генома в проекциях пространства трех первых главных компонент для Stigeoclonium helveticum

На рис. 1, а, видно, что точки сформированы в трехлучевую структуру, которая кластеризуется относительно кодирующих и некодирующих областей генома. В центральном кластере расположены точки, соответствующие некодирующим областям, они отмечены коричневым цветом. Точки кодирующих областей распределены по лучам следующим образом: первому лучу соответствуют фазы B_2 и F_2 (точки светло-желтого и темно-желтого цветов), второму лучу соответствуют фазы B_0 и F_1 (точки темно-сиреневого и светло-зеленого цветов), и третьему лучу соответствуют фазы B_1 и F_0 (точки светло-сиреневого и темно-зеленого цветов). На рис. 1, б, видно, что кроме трехлучевой структуры выделяется кластер, изолированный от остальных точек.

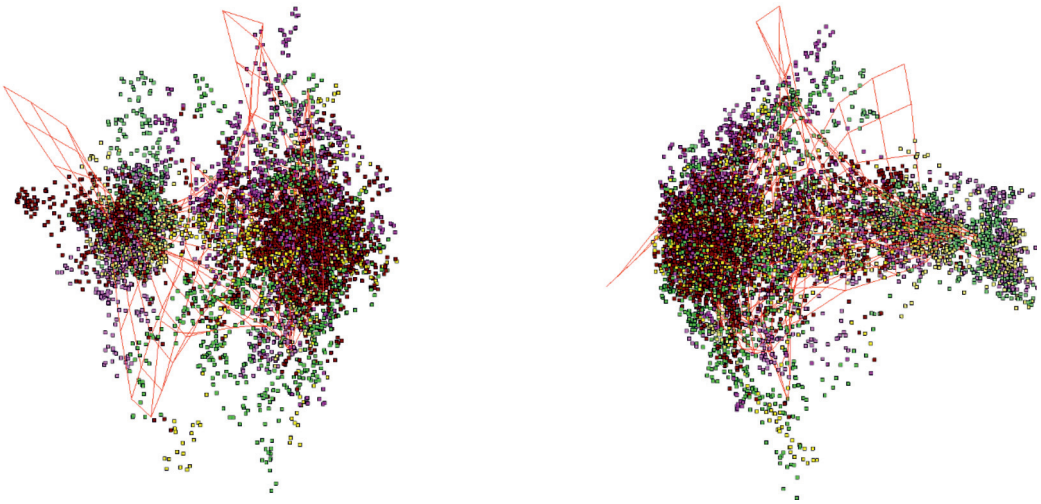
Кроме этого, были геномы, чья структура отличалась от характерной. У геномов хлоропластов *Oltmannsiellopsis viridis, Stigeoclonium helveticum, Cyanidioschyzon merolae, Chromera velia, Cyanidiaceae sp. MX-AZ01, Xylochloris irregularis и Aureococcus anophagefferens* отсутствует кластер, изолированный от остальных точек (см. рис. 2).

Выделилась группа водорослей, у которых наблюдается шестилучевая структура. К этой группе относятся: *Klebsormidium flaccidum, Chlorella vulgaris, Micromonas commoda, Chlorella sorokiniana, Chlorella sp. ArM0029B, Chlorella variabilis, Prasinoderma colonialis, Phaeocystis globosa.* На рис. 3 показана структура генома, характерная для этой группы хлоропластов.



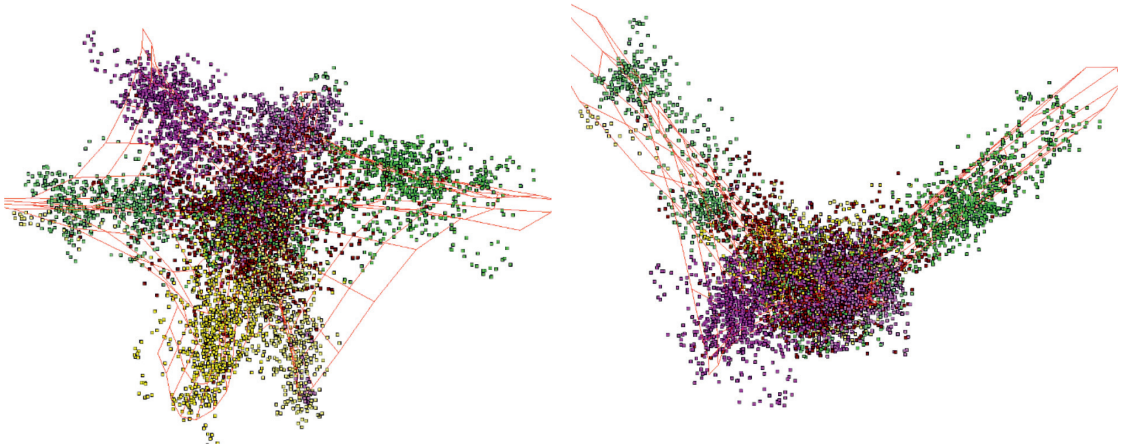
«анфас» *«в профиль»*

Рис. 3. Шестилучевая структура генома (Prasinoderma colonialis)



«анфас» *«в профиль»*

Рис. 4. Структура генома Euglena longa в проекциях пространства трех первых главных компонент



«анфас» *«в профиль»*

Рис. 5. Структура генома Ostreococcus tauri в проекциях пространства трех первых главных компонент

Кроме того, у трех видов водорослей была обнаружена двухъядерная структура. Это *Euglena longa, Euglenaformis proxima и Monomorphina aenigmatica*. Они все относятся к семейству *Euglenaceae.* На рис. 4 показана структура генома для *Euglena longa*.

У *Ostreococcus tauri* наблюдается шестилучевая структура и отсутствует кластер, изолированный от остальных точек (рис. 5).

Заключение

Таким образом, было установлено, что для большинства геномов хлоропластов одноклеточных и многоклеточных водорослей структура генома в пространстве первых трех главных компонент очень похожа: выделяется центральный кластер, состоящий из участков некодирующих областей, и три луча, состоящих из участков кодирующих областей. Кроме того, присутствует изолированная группа точек, не входящая в трехлучевую структуру. В [14] была рассмотрена структура геномов хлоропластов наземных растений. Хочется отметить, что структура геномов хлоропластов наземных растений гораздо более однородна: отсутствие изолированного кластера наблюдается только у двух видов. Шестилучевой структуры у геномов хлоропластов наземных растений обнаружено не было. Больший разброс по структуре геномов хлоропластов для водорослей по сравнению с наземными растениями может объясняться гораздо большими различиями в среде обитания у водорослей. В [1, 2] было показано, что структура геномов бактерий определяется их *GC*-составом. У бактерий шести лучевая структура наблюдалась при *GC*-составе больше 0,6. У водорослей шестилучевая структура была обнаружена при значениях *GC*-состава от 0,31 до 0,42. Причем у других видов водорослей при тех же значениях *GC*-состава структура генома была трехлучевой.

Список литературы

1. Gorban A.N., Zinovyev A.Yu., Popova T.G. Seven clusters in genomic triplet distributions // Silico Biology. – 2003. – vol. 3. no. 4. – P. 471–482.

2. Gorban A.N., Zinovyev A.Yu., Popova T.G. Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences // Silico Biology. – 2005. – vol. 5. no. 3. – P. 265–282.

3. Martin W.F., Roettger M., Kloesges T., Thiergart T. Woehle C., Gould S.B., Dagan T. Modern endosymbiotic theory: getting lateral gene transfer into the equation // J. Endocyt Cell Res. – 2012. – vol. 23. – P. 1–5.

4. Elias M., Archibald J.M. Sizing up the genomic footprint of endosymbiosis // Bioessays. – 2009. – vol. 31. – P. 1273–1279.

5. Waseemuddin M.D. Genomic Studies in Bacteria, Mitochondria and Chloroplast in Relation with Endo-Symbiotic Theory // International Journal of Scientific and Research Publications. – 2008. – vol. 6, no. 3. – P. 64–65.

6. Nakayama T., Archibald J.M. Evolving a photosynthetic organelle // BMC Biol. – 2012. – vol. 10. – P. 35–38.

7. Keeling P.J., Archibald J.M. Organelle evolution: What's in a name? // Curr. Biol. – 2008. – vol. 18, no. 8. – P. 345–347.

8. McFadden G.I. Origin and Evolution of Plastids and Photosynthesis in Eukaryotes // Cold Spring Harb Perspect Biol., 2014, vol. 6, no. 4: a016105.

9. Stiller J.W. Plastid endosymbiosis, genome evolution and the origin of green plants // Trends Plant Sci. – 2007. – vol. 12. – P. 391–396.

10. Ku C., Nelson-Sathi S., Roettger M., Garg S., Hazkani-Covo E., Martin W.F. Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes // Proc Natl Acad Sci USA. – 2015. – vol. 112, no. 33. – P. 10139–10146.

11. Stiller J.W. Plastid endosymbiosis, genome evolution and the origin of green plants // Trends Plant Sci. – 2007. – vol. 12. – P. 391–396.

12. Сенашова М.Ю., Садовский М.Г. Пространственная структура геномов цианобактерий // Международный журнал фундаментальных и прикладных исследований. – 2017. – № 11–2. – С. 255–259.

13. Зиновьев А.Ю. Программа визуализации данных VidaExpert. URL: http://bioinfo-out.curie.fr/projects/vidaexpert/ (дата обращения: 19.12.17).

14. Сенашова М.Ю., Садовский М.Г. Семикластерная структура геномов хлоропластов отражает филогению их носителей // Международный журнал прикладных и фундаментальных исследований. – 2016. – № 12–7. – С. 1167–1173.