УДК 573.22:575.89

ПРОСТРАНСТВЕННАЯ СТРУКТУРА ГЕНОМОВ ЦИАНОБАКТЕРИЙ

[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

Представлены результаты, полученные при изучении пространственной структуры геномов цианобактерий. В качестве структуры в нашей работе понимается расположение в пространстве частот триплетов точек, соответствующих выделенным участкам генома цианобактерий. Для каждого участка длины Δ со сдвигом t вычислялся частотный словарь троек символов без пересечений. Частоты рассматривались как координаты в 64-мерном пространстве. Таким образом, каждому участку генома сопоставлялась точка в пространстве частот. Было проанализировано 7 геномов цианобактерий, размещенных в EMBL-банке. Одна координата (с минимальным стандартным отклонением) отбрасывалась и в дальнейшем рассматривалось 63-мерное пространство частот. Для визуализации полученного множества точек была использована программа VidaExpert. С ее помощью для каждого генома были построены проекции в пространство первых трёх главных компонент из 63-мерного пространства частот. Мы обнаружили, что геномы цианобактерий обладают одинаковой структурой, представляющей собой своеобразный клубок из нитей. Причем нити образованы точками, соответствующими последовательным участкам генома.

Ключевые слова: геном, триплет, частота, структура данных

SPATIAL STRUCTURE OF GENOMES OF CYANOBACTERIA

[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: msad@icm.krasn.ru;

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

The results obtained in the study of the structure of the genomes of cyanobacteria are presented. By structure in our work is meant the location in space of triplet frequencies of points corresponding to fragments of the genome of cyanobacteria. For each length-shifted fragments, a frequency dictionary of symbol triples without intersections was computed. Frequencies were considered as coordinates in a 64-dimensional space. Thus, each point of the genome was compared with a point in the frequency space. Seven genomes of cyanobacteria located in the EMBL-bank were analyzed. One coordinate (with a minimal standard deviation) was discarded and a 63-dimensional frequency space was subsequently considered. For each genome in space of frequencies by means of program VidaExpert projections from 63-dimensional space in space of first three main components have been constructed. This allowed us to visualize the structure of genomes. It was found that the genomes of cyanobacteria have the same structure, which is a kind of tangle of filaments. And the threads are formed by points corresponding to consecutive fragments of the genome.

Keywords: genom, triplet, frequency, data pattern

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 всё ещё далека от завершения. Более того, исследователи выявляют всё новые и новые структурные элементы либо новые виды и формы взаимодействий и взаимоотношений между структурными элементами биологических макромолекул, а развитие техники и инструментов исследований лишь усугубляет эту ситуацию.

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 мате-

риала, который берётся 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 удаётся 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. Своего рода расплатой за такое удобство является заметная трудность в определении филогении бактерий, особенно для таксонов высокого уровня.

Исследование структур в генетических последовательностях также является важной задачей, осложняющейся чрезвычайно большим разнообразием структур, которые можно найти и выделить в молекулах ДНК, даже если не обращать внимания на их химические свойства. Настоящая работа посвящена изучению структур в геномах цианобактерий.

Под структурой мы будем понимать различие (либо подобие) статистических свойств отдельных формально выделяемых участков генома на уровне триплетов. Иными словами, структура, рассматриваемая в этой работе, – это взаимное расположение различных (формально выделяемых) участков генома сравнительно небольшой длины в пространстве частот триплетов, которые подсчитываются в пределах указанных участков; подробности изложены в разделе «Методы исследования».

Такой подход к изучению связи структуры геномов с их GC-составом был впервые предложен в работах Горбаня с соавторами [1, 2]. Этот же подход (с небольшой модификацией) используется и нами; один из мотивов использования метода, предложенного в указанных работах, – теория симбиогенеза [3–9]: поскольку согласно этой теории современные хлоропласты и цианобактерии имеют общего предка. Если эта теория верна, то можно надеяться найти какие-то признаки подобия структур, выделяемых как в бактериальных геномах, так и в геномах хлоропластов. Следует сказать сразу, что были обнаружены существенные различия, а не подобие. Из этого не следует, что теория происхождения хлоропластов от цианобактерий неверна; это означает, что в процессе эволюции этих двух генетических изолированных систем произошла сильная дивергенция.

Материалы и методы исследования

Введём основные понятия. 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. Для этой последовательности символов мы будем составлять частотный словарь толщины 3. Под частотным словарем W_3 толщины 3 символьной последовательности, соответствующей ДНК, будем понимать список всех триплетов $v_1v_2v_3$ идущих подряд символов с указанием частот этих триплетов. Таких комбинаций может быть 64. В качестве частоты f_ω рассматривается отношение количества копий n_ω выбранного триплета к общему количеству всех триплетов N, где N – сумма всех n_ω:

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

Любой частотный словарь W_3 ставит в соответствие геному множество точек в 64-мерном метрическом пространстве. Для оценки близости двух геномов используется метрика Евклидового пространства, определяющая расстояние между двумя точками:

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

Для исключения линейной связи между частотами триплетов (поскольку сумма частот равна единице) один из триплетов удалялся из рассмотрения. Это позволяет уменьшить погрешность, вносимую линейной зависимостью при обработке данных статистическими методами, например при корреляционном анализе или при использовании метода главных компонент. Вообще говоря, из рассмотрения можно удалять какой угодно триплет. Тем не менее есть ряд подходов для выбора удаляемого триплета. Один из подходов в качестве удаляемого триплета предлагает выбирать триплет с максимальной частотой. Тем более если значение частоты для исключаемого триплета существенно больше (например, на порядок) значения частоты триплета, идущего за ним.

Так же часто используется подход, при котором удаляется триплет с минимальным стандартным отклонением, вычисленным по всей совокупности частотных словарей данного генома. Такой выбор обусловлен тем, что вклад этого триплета в разделение точек в пространстве минимален. В случае равенства стандартного отклонения 0 различий по этому триплету не наблюдается. Мы использовали именно этот подход. Размерность пространства уменьшается на единицу и становится 63-мерным. В рассмотренных нами геномах в большинстве случаев удалялись триплеты GCG и CGC.

Для обнаружения структуры в генетической последовательности проводилась предварительная обработка, которая ставила в соответствие данной

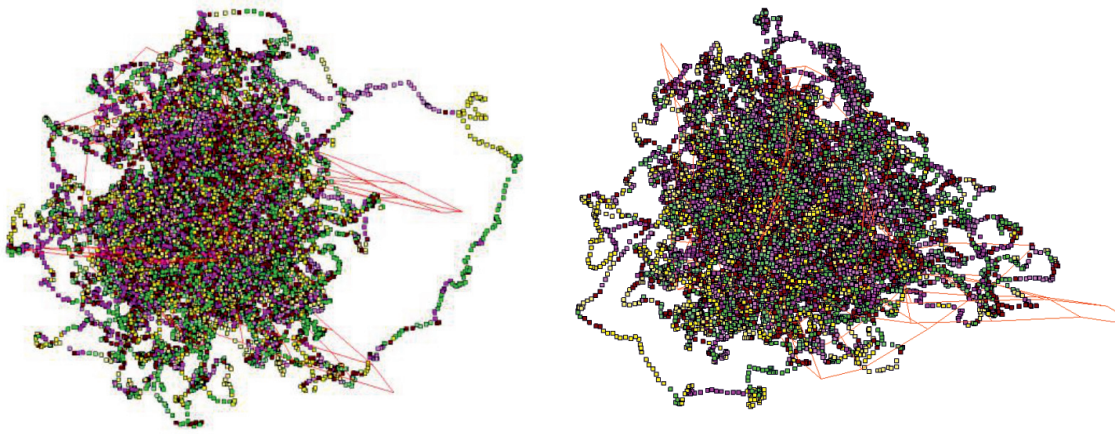
последовательности множество точек в 63-мерном пространстве триплетов. Делалось это следующим образом: последовательность сканировалась окном длины Δ с шагом t. Для каждого положения i рамки определялся участок генетической последовательности, совпадающий с рамкой считывания, для которого вычислялся частотный словарь $W_3^{(i)}$, соотносящийся с i-ой точкой 64-мерного пространства. Кроме того, с каждой точкой 64-мерного пространства 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 тем, попал рассматриваемый участок в кодирующую или некодирующую область последовательности. Участок относится к кодирующим, если он целиком попадал в кодирующую область последовательности. Для некодирующего участка соответствующая ему точка помечается символом J. Если участок относится к кодирующим, для него возможны 6 вариантов маркировки: B_0, B_1, B_2, F_0, F_1, F_2. Для кодирующего участка, аннотированного в последовательности как считывающийся в прямом направлении, вычисляется 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 символом B_0, B_1 или B_2. Для участка, аннотированного как считывающийся в обратном направлении, вычисляется 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 Δ = 6003, шаг t = 101.

Для полученного множества точек в программе *VidaExpert* [10] вычислялась и визуализировалась проекция в пространство первых трёх главных компонент из 63-мерного пространства. Для получения двумерных изображений строились проекции на плоскость первых двух главных компонент и второй и третьей главной компоненты. Группы точек, в зависимости от принадлежности участка, были обозначены разным цветом. Для некодирующих участков точки изображены на рисунках коричневым цветом, для участков, соответствующих относительным фазам B_0 и F_0, точки изображены темно-малиновым и светло-малиновым цветом, для участков, соответствующих относительным фазам B_1 и F_1, точки изображены темно-зеленым и светло-зеленым цветом, а для соответствующих относительным фазам B_2 и F_2 точки изображены темно-желтым и светло-желтым цветом.

Результаты исследования и их обсуждение

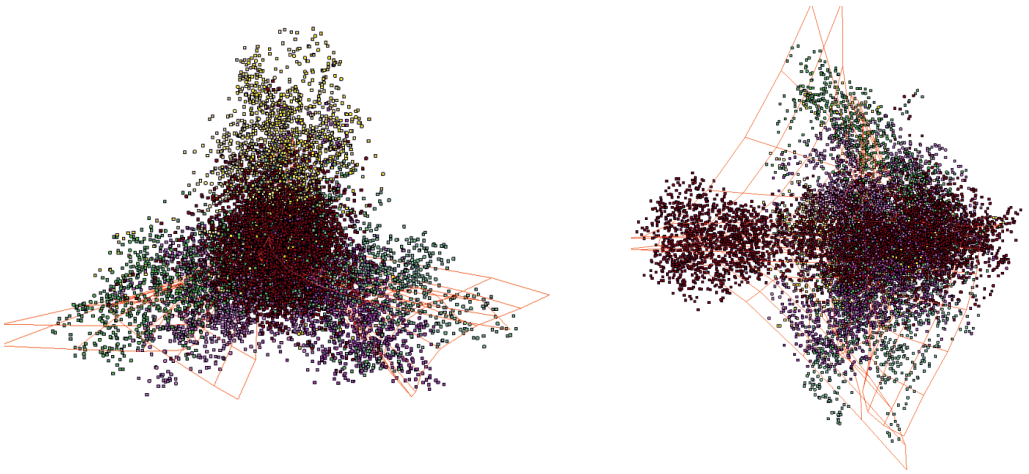
Гипотеза о происхождении хлоропластов от одноклеточных свободноживущих фотосинтезирующих бактерий позволяет ожидать, что структура геномов цианобактерий будет подобна аналогичной структуре хлоропластов. Мы рассмотрели структуры для геномов цианобактерий, депонированных в EMBL-банке, а именно *Microcystis aeruginosa NIES-843, Nostoc sp. PCC 7107, Pleurocapsa sp. PCC 7327, Chroococcidiopsis thermalis PCC 7203, Gloeocapsa sp. PCC 7428, Anabaena cylindrica PCC 7122, Synechocystis sp. PCC 6714.* Эти цианобактерии относятся к трем порядкам: Chroococcales, Nostocales, Pleurocapsales.



а) *б)*

Рис. 1. Структура данных Microcystis aeruginosa (а) и Nostoc sp. PCC 7107 (б) в проекции на плоскость первых двух главных компонент

На рис. 1 показан характерный вид структуры геномов цианобактерий в проекции на плоскость первых двух главных компонент. Как видно из рис. 1, структура геномов цианобактерий представляет собой своеобразные клубки из цепочек, точки в которых расположены в той же последовательности, в которой соответствующие участки расположены в геноме.



а) Проекция в плоскость первых двух главных компонент

б) Проекция в плоскость второй и третьей главных компонент

Рис. 2. Типичный вид распределения участков хлоропластных геномов наземных растений по частотам троек нуклеотидов в проекциях пространства первых трех главных компонент (приведена структура генома Lolium perenne)

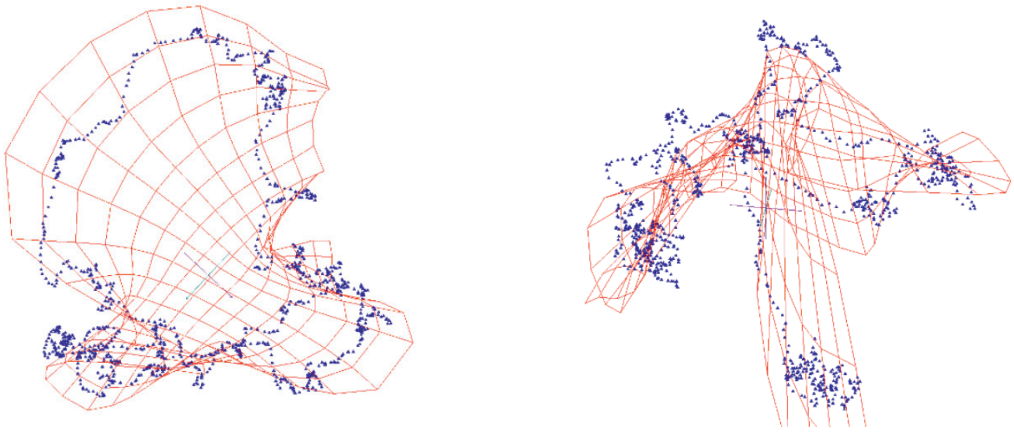


Рис. 3. Структура геномов Microcystis aeruginosa и Nostoc при шаге t, кратном трем

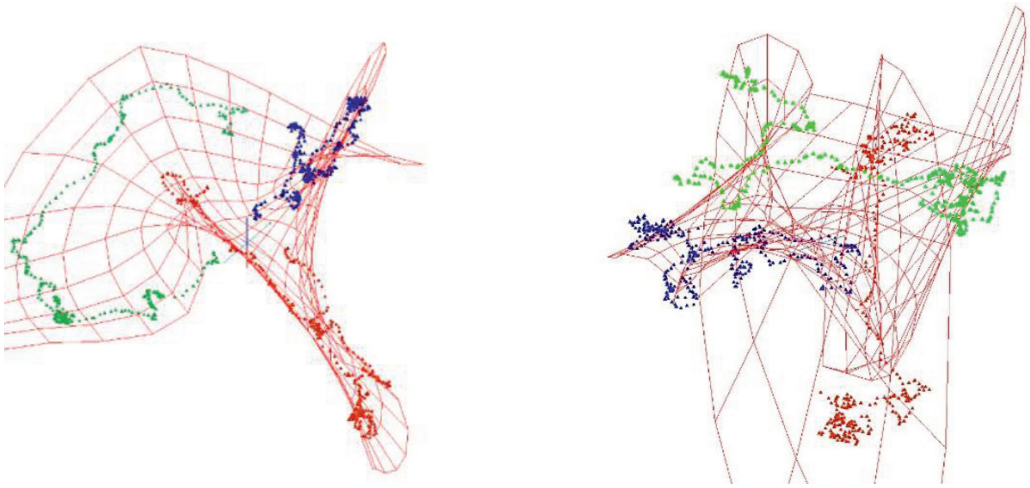


Рис. 4. Структура геномов Microcystis aeruginosa и Nostoc при шаге t, не кратном трем

Структура геномов хлоропластов наземных растений была рассмотрена в [11]. Исследование показало, что подавляющее большинство геномов хлоропластов имеет восьмикластерную структуру (рис. 2). Семь кластеров составляют трехлучевую структуру – центральное ядро из точек, соответствующих некодирующим участкам и лучи, соответствующие относительным фазам. Первый луч включает точки с маркировкой B_0, F_1, второй – с маркировкой B_1, F_0 и третий – с маркировкой B_2, F_2 (рис. 2, а). И есть еще выделенный кластер, который видно на рис. 2, б. Как видно из рис. 1, 2 структура геномов цианобактерий существенно отличается от структуры геномов хлоропластов. Это наблюдение может свидетельствовать о том, что расхождение от одноклеточных свободноживущих фотосинтезирующих бактерий современных цианобактерий и хлоропластов произошло очень давно и о независимом характере их развития.

Кроме того, было обнаружено, что «клубок», соответствующий структуре цианобактерий в пространстве главных компонент, состоит из одной или из трех нитей, в зависимости от величины шага t. В случае кратности шага трем клубок состоит из одной нити, в противном случае нитей три. Причем точки в одной нити следуют друг за другом последовательно, как и соответствующие им участки генома. В случае трех нитей одна нить состоит из точек, номер участков в геноме которых делится на 3 нацело, вторая нить состоит из точек, номер участков для которых делится на 3 с остатком 1, третья нить – из точек, номер участков для которых делится на 3 с остатком 2. На рис. 3 показаны участки геномов *Microcystis aeruginosa и Nostoc*, включающих первую тысячу точек в пространстве первых трех главных компонент для $\Delta = 6003$ и $t = 303$. На рисунке видно, что структура генома представляет собой одну нить. На рис. 4 показаны участки этих же геномов для $\Delta = 6003$ и $t = 91$.

Выводы

Структура геномов цианобактерий существенно отличается от структуры геномов хлоропластов. Кроме того, структура геномов цианобактерий существенно отличается и от структуры геномов иных бактерий. Семикластерная структура генома у цианобактерий отсутствует. Это также выделяет цианобактерии среди других бактерий и подчеркивает их особенности, первой среди которых выделяется способность к фотосинтезу. В частности, это наблюдение может свидетельствовать об очень древнем расхождении современных цианобактерий и хлоропластов от общего предка и о весьма сложных путях их эволюции.

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. Zimorski V., Ku Ch., Martin W.F., Gould S.B. Endosymbiotic theory for organelle origins // Curr. Opin. Microbiol. –2014. –vol. 22. –P. 38–48.

4. Falcon L.I., Magallon S., Castillo A. Dating the cyanobacterial ancestor of the chloroplast // ISME J. – 2010. – vol. 4. – P. 777–783.

5. Chan Ch. X., Bhattacharya D. Plastid Origin and Evolution. // eLS. John Wiley & Sons, Ltd: Chichester, 2011.

6. Kleine T., Maier U.G., Leister D. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis // Annu. Rev. Plant. Biol. –2009. –vol. 60. –P. 115–138.

7. Howe C.J., Barbrook A.C., Nisbet R.E.R., Lockhart P.J., Larkum A.W.D. The origin of plastids // Philos. Trans. R. Soc. Lond. B: Biol. Sci. –2008. – vol. 363. – P.2678–2685.

8. Lane C.E., Archibald J.M. The eukaryotic tree of life: endosymbiosis takes its TOL // Trends Ecol. Evol. – 2008. – vol. 23. –P. 268–275.

9. Moustafa A., Beszteri B., Maier U.G., Bowler C., Valentin K., Bhattacharya D. Genomic footprints of a cryptic plastid endosymbiosis in diatoms // Science. – 2009. – vol. 324. – P. 1724–1726.

10. Зиновьев А.Ю. Программа визуализации данных VidaExpert // сайт. – URL: http://bioinfo-out.curie.fr/projects/vidaexpert/ (дата обращения: 19.09.17).

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