ЖУРНАЛ ОБЩЕЙ БИОЛОГИИ, 2018, том 79, № 2, с. 124–134

УДК 57.015 + 573.2

ВОСЬМИКЛАСТЕРНАЯ СТРУКТУРА ГЕНОМОВ ХЛОРОПЛАСТОВ НАЗЕМНЫХ РАСТЕНИЙ

© 2018 г. М. Г. Садовский[1,2], М. Ю. Сенашова[1], А. В. Малышев[1]

[1]Институт вычислительного моделирования СО РАН
660036, Красноярск, Академгородок
e-mail: msad@icm.krasn.ru
[2]Сибирский Федеральный университет
Институт фундаментальной биологии и биотехнологии
660041, Красноярск, просп. Свободный, 79

Поступила в редакцию 03.04.2017 г.

В данной работе было проверено, является ли семикластерная структура, обнаруженная у бактерий, универсальной и наблюдаются ли все обнаруженные свойства такой структуры для других геномов. Исследована внутренняя структурированность геномов хлоропластов и цианобактерий, показано, что эта структурированность существенно отличается от ранее обнаруженной для бактериальных геномов. Под структурированностью понимается кластеризация частотных словарей триплетов отдельных фрагментов генома, определяемых регулярным порядком, вне зависимости от функциональной роли того или иного участка. Кластеризация проводилась методом упругих карт.

Изучение особенностей и деталей структуры нуклеотидных последовательностей является важнейшей задачей современной биологии. Такого рода исследования ведутся в двух аспектах – структурно-функциональном и эволюционном. Выявление связи структуры и функции представляет собой классическую проблему молекулярной и системной биологии и, несмотря на обширный поток публикаций и работ в данном 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, а развитие техники и инструментов исследований лишь усугубляет эту ситуацию.

Эволюционный аспект таких исследований вполне понятен: изучая особенности структуры биологических макромолекул (ДНК либо белковых последовательностей) у разных организмов можно точнее составить картину того, как эволюционировали те или иные биологические системы – от конкретных видов (рас, штаммов и т.п.) до экосистем и глобальных сообществ.

Такого рода исследования всегда затруднены выбором и качеством того биологического материала, который берется в рассмотрение. Дело даже не в ошибках (к примеру) секвенирования и/или аннотирования генетических последовательностей, неизбежных во многих случаях, а в большой сложности таких объектов, как геномы либо отдельные хромосомы. Фактически приходится анализировать триаду характеристик: структуру, функцию и филогению. Эти элементы триады плотно переплетены друг с другом и оказывают друг на друга существенное влияние, которое 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. Своего рода расплатой за такое удобство является заметная трудность в определении филогении бактерий, особенно для таксонов высокого уровня. Еще более простым и контролируемым объектом для подобных исследований являются геномы органелл – хлоропластов и митохондрий, поскольку для них полностью исключается влияние различий в кодируемых функциях: в пределах одной группы органелл функциональные различия отсутствуют.

Настоящая работа посвящена изучению связи структуры нуклеотидных последовательностей и таксономии их носителей на примере геномов хлоропластов. Поток работ, посвященных

изучению эволюционных процессов и отношений в царстве растений на основе анализа особенностей структуры последовательностей геномов (Neale et al., 1988; Leliaert et al., 2012; Carbonell-Caballero et al., 2015) либо отдельных фрагментов и генов хлоропластов (Gielly, Taberlet, 1994; Katayama, Ogihara, 1996; Milanowski et al., 2001; Shaw et al., 2005; Marazzi et al., 2006; Dong et al., 2012), велик. При этом часто утверждения об эволюции вида делаются на основании анализа изменений генома хлоропластов.

Исследования структур в генетических последовательностях также является важной задачей, осложняющейся большим разнообразием структур, которые можно найти и выделить в молекулах ДНК, даже если не обращать внимания на их химические свойства. Под структурой мы будем понимать различие (либо подобие) статистических свойств отдельных формально выделяемых фрагментов генома на уровне триплетов. Иными словами, структура в рамках настоящей работы – это взаимное расположение различных (формально выделяемых) фрагментов генома сравнительно небольшой длины в пространстве частот триплетов, которые подсчитываются в пределах указанных фрагментов; подробности представлены в разделе “Материалы и методы”.

Данный подход к изучению структурированности геномов и их связи с их GC-составом был впервые предложен в работах Горбаня с соавт. (Gorban et al., 2003a, b, 2005). Этот же подход (с небольшой модификацией) используется и нами; один из мотивов использования метода, предложенного в указанных работах – теория симбиогенеза (Mereschkovsky, 1905; Мережковский, 1909): согласно этой теории, современные хлоропласты и цианобактерии имеют общего предка. Если эта теория верна, то можно надеяться найти какие-то признаки подобия структур, выделяемых в бактериальных геномах и в геномах хлоропластов. Анонсируя основной результат, скажем сразу, что были обнаружены существенные различия, а не подобие. Из этого не следует, что теория симбиотического происхождения хлоропластов неверна, из этого следует, что в ходе эволюции этих двух генетических изолированных систем произошла сильная дивергенция.

МАТЕРИАЛЫ И МЕТОДЫ

Введем основные понятия. Мы будем рассматривать генетическую последовательность длины L, состоящую из символов алфавита $\aleph = \{A, C, G, T\}$. Если последовательность содержит символы, отличающиеся от символов алфавита $\aleph$, то такие символы из последовательности удаляются, а длина последовательности уменьшается на число таких символов. Поскольку удаление части символов может влиять на положение рамки считывания во фрагментах, выделяемых далее по последовательности, поясним детали процедуры. Необходимо сказать, что из всех проанализированных геномов лишь у пяти были обнаружены символы, не входящие в алфавит. В тех последовательностях, в которых такие символы были обнаружены, их удаление (а они всегда шли короткими группами) сопровождалось пересчетом положения всех последующих нуклеотидов, что сохраняло относительное положение рамки считывания в таких фрагментах. Для этой последовательности мы будем составлять частотный словарь толщины 3. Частотный словарь W_3 толщины 3 символьной последовательности, соответствующей ДНК – это список всех троек $v_1v_2v_3$ 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 $k = 1$. В работах Горбаня с соавт. (Gorban et al., 2003a, b, 2005) использовался 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. Такой выбор словаря обусловлен тем, что одной из целей настоящей работы было сравнение результатов, полученных нами и в работах Горбаня с соавт. (Gorban et al., 2003a, b, 2005). Частота f_ω – это отношение числа копий n_ω данного слова к общему числу всех триплетов N, где N – сумма всех n_ω:

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

Всякий частотный словарь W_3^3 отображает геном в 64-мерное метрическое пространство; близость двух геномов задается естественным образом – например, как близость двух точек в Евклидовой метрике:

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

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.). Формально исключить можно любой триплет, однако есть несколько эвристических подходов к исключению. Первый подход состоит в том, чтобы исключить самый большой по значению частоты триплет, особенно если значение его частоты на порядок (или около того) превосходит значение частоты следующего за ним (по этой величине) триплета.

Второй подход состоит в том, чтобы 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). Мы использовали второй подход. Таким образом, рассматриваемое нами пространство точек становится 63-мерным. В наших исследованиях в основном исключались триплеты GCG и CGC, подробности см. в разделе “Результаты”.

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 окном длины Δ с шагом t. Для каждого положения i рамки определялся участок генетической последовательности, совпадающий с рамкой считывания, для которого вычислялся частотный словарь $W_3^{(i)}$, соответствующий i-й точке в 63-мерном пространстве. Кроме того, с каждой точкой в 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 в зависимости от 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 шесть вариантов маркировки: B_0, B_1, B_2, F_0, F_1, F_2. Если кодирующий участок в последовательности 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 символом F_0, F_1 или F_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 B_0, B_1 или B_2.

Данные для исследования брались в базе EMBL-банка. Были отобраны 185 хлоропластов наземных растений. Для всех генетических последовательностей длина рамки считывания $\Delta = 603$, шаг $t = 11$. Такой выбор рамки считывания обусловлен следующими соображениями: во-первых, такая длина фрагмента близка к характерной длине гена; во-вторых, уменьшение рамки считывания хотя бы вдвое приводит к значительному вырождению частотных словарей. Это означает, что значительное (до трети) число всех слов в частотном словаре, соответствующем выделенному фрагменту, имеют нулевые частоты либо представлены в одной копии; тем самым, различимость фрагментов по этим триплетам отсутствует. В-третьих, увеличение длины выделяемого фрагмента до $\Delta \approx 1500$ не меняет качественно картину, в то время как увеличение Δ до 3000 и выше нуклеотидов приводит к образованию большого числа четко отделенных друг от друга плотных кластеров, в которых представлены фрагменты, относящиеся к разным относительным фазам. Косвенно это свидетельствует о том, что фрагменты длиной 3000 нуклеотидов и выше почти всегда содержат две кодирующие области.

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: вид данных “анфас” и “в профиль”. Точки данных были раскрашены в соответствии с их относительной фазой. Точки, соответствующие не кодирующим участкам, изображены на рисунках серыми кругами, точки, соответствующие относительным фазам F_0 и B_0, изображены светло-серыми и темно-серыми квадратами, точки, соответствующие относительным фазам F_1 и B_1, изображены светло-серыми и темно-серыми треугольниками, а точки, соответствующие относительным фазам F_2 и B_2, изображены светло-серыми и темно-серыми ромбами. Кроме того, для каждого генома вычислялся GC-состав, т.е.

отношение суммарного числа нуклеотидов C и G к общему числу нуклеотидов в геноме.

Для облегчения восприятия результатов введем точные определения терминов, которые будут использоваться ниже. Выделяемые рамкой считывания длины Δ фрагменты помечались тремя разными метками: *ген*, *джанк* и *хвост*. При этом исходно фрагменты получали одну из двух меток: *ген* или *джанк*. Следует подчеркнуть, что фрагменты выделяются только из одной (прямой) цепи ДНК и каждый такой фрагмент может содержать 1) кодирующую область из прямой цепочки с тремя возможными сдвигами относительно первой позиции фрагмента; 2) кодирующую область из обратной цепочки с тремя возможными рамками считывания (определяемые по "зеркальному отражению"); 3) некодирующие области; и, наконец, 4) сочетание некодирующих и кодирующих областей в разных пропорциях. Эти последние фрагменты вносят шум в распределение, но их относительная доля не высока. Понятно, что выбор порога доли кодирующей области в выделяемом фрагменте будет оказывать некоторое влияние на результаты распределения; в рамках настоящей работы использовался весьма жесткий критерий: в категорию "ген" попадали лишь те фрагменты, которые целиком лежали в кодирующей области. Это обусловлено тем, что в противном случае выбор доли кодирующих и некодирующих участков во фрагментах, которые будут считаться "геном" очень субъективен. Влияние величины доли некодирующих областей в выделяемых фрагментах нами не исследовалось.

Фрагмент с номером *i* получал метку *ген*, если он попадал в ту часть генома, которая в аннотации, содержащейся в голове файла EMBL–банка, была помечена как CDS, либо как exon, либо как rRNA, если длина такового превышала длину фрагмента. При этом фрагмент получал такую метку лишь при условии полного попадания в участок, аннотированный как описано выше.

Фрагмент с номером *i* получал метку *джанк*, если он не получал метку *ген*, при условиях, описанных выше.

Фрагмент с номером *i* получал метку *хвост* не на основании данных аннотирования генома, но на основании результатов кластеризации в пространстве главных компонент. Метку *хвост* получали те фрагменты, которые в пространстве трех первых главных компонент образовывали четко выраженный сравнительно небольшой кластер, лежащий на оси симметрии получающейся структуры. В *хвост* попадают фрагменты, относящиеся к *джанку* и rRNA. Следует сказать, что некоторые фрагменты, физически располагающиеся в хвосте, были помечены как джанк, несмотря на то, что они содержали гены tRNA, которые существенно короче выделяемого фрагмента.

РЕЗУЛЬТАТЫ

Начнем представление результатов с описания базы исследованных геномов. Для каждого генома выделялись три типа фрагментов: *ген*, *джанк* и *хвост* (их детальное определение см. выше). Кроме того, мы определяли GC-состав для:

а) всего генома в целом,

б) только для фрагментов, помеченных как *ген* (вне зависимости от того, в какой ориентации они находятся),

в) для фрагментов, помеченных как *джанк* и

г) для фрагментов, располагающихся в *хвосте* (см. ниже).

На рис. 1 показано распределение геномов по значениям GC-состава: точнее, на рисунке представлено 185 геномов хлоропластов, которые упорядочены по возрастанию значений GC-состава, определяемого по всему геному в целом. На этом множестве геномов выделяется очень обширная группа, у которой значение GC-состава лежит в интервале (0.35; 0.40).

У хлоропластов *Orthotrichum rogeri*, *Syntrichia ruralis*, *Physcomitrella patens*, *Marchantia polymorpha*, *Sanionia uncinata*, *Anthoceros angustus*, *Ptilidium pulcherrimum*, *Equisetum arvense*, *Glycyrrhiza glabra*, *Trifolium subterraneum*, *Orobanche gracilis*, *Taxus mairei*, *Millettia pinnata*, *Pisum sativum*, *Juniperus virginiana* и *Juniperus bermudiana* значения GC-состава лежат ниже отметки в 0.35 (список приведен в порядке возрастания значений GC-состава). Противоположную группу составляют геномы видов *Aneura mirabilis*, *Lygodium japonicum*, *Pteridium aquilinum*, *Ophioglossum californicum*, *Marsilea crenata* и *Myriopteris lindheimeri*, у которых значения GC-состава превышают уровень 0.40 (список

Коэффициенты корреляции между значениями GC-состава для разных фрагментов и геномов хлоропластов в целом

Тип фрагмента	Ген	Джанк	Хвост
Геном	0.9745	0.9617	0.6248
Ген		0.9218	0.6285
Джанк			0.5939

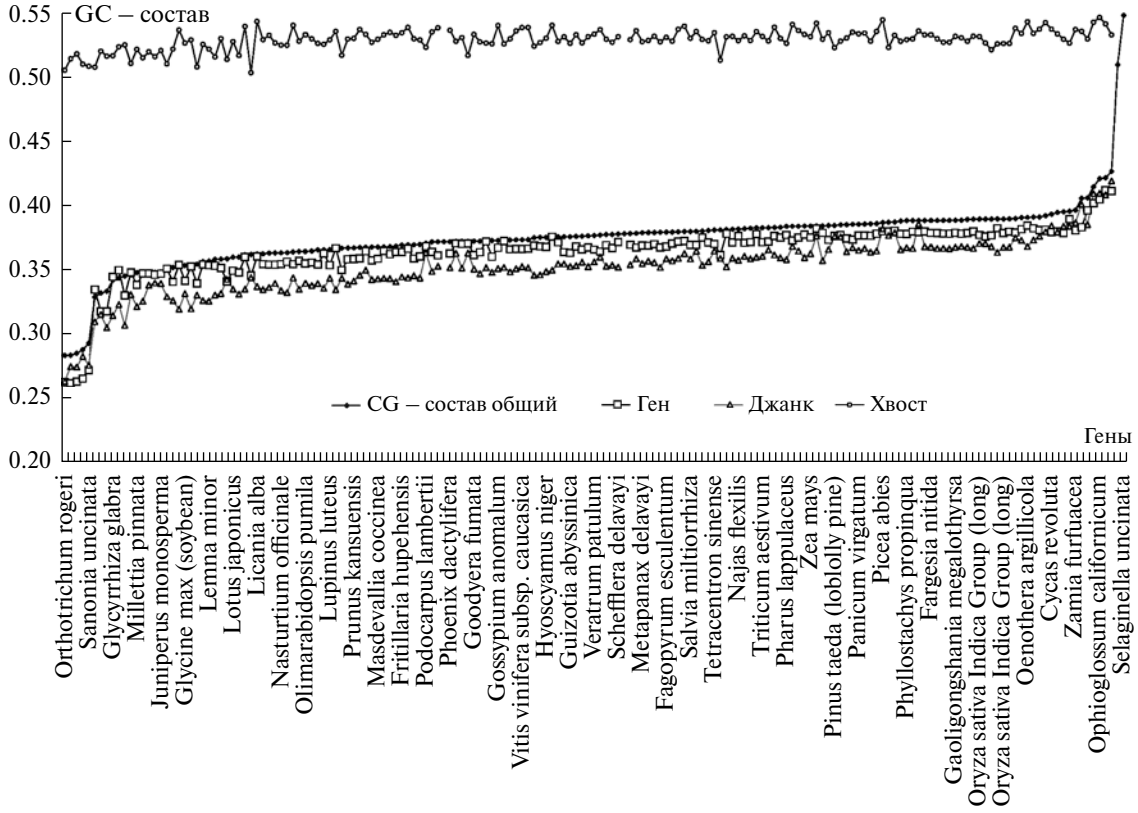

Рис. 1. GC-состав хлоропластных геномов и их фрагментов. Виды ранжированы по GC-составу хлоропластных геномов в целом.

также приведен в порядке возрастания значений); наконец, есть два генома-рекордсмена – *Selaginella moellendorffii* и *S. uncinata*, у которых значение GC-состава составляет 0.51003 и 0.54849 соответственно.

Данные (рис. 1) показывают, что значения GC-состава для фрагментов геномов, содержащих как гены, так и джанк, в целом близки к значениям GC-состава, определяемого для всего генома. Тот факт, что согласованность в значениях GC-состава для генов и джанка со значениями GC-состава для всего генома весьма высока, следует из таблицы, в которой представлены коэффициенты корреляции между значениями GC-состава для всего генома и для фрагментов, содержащих гены и джанк. Напротив, GC-состав фрагментов геномов, которые составляют "хвост" в распределении фрагментов по кластерам (см. ниже), большее 0.50 (рис. 1, верхняя кривая). Два разрыва в верхней кривой на рисунке указывают на пробелы в данных; это не ошибка расчетов, это недостаток аннотации хлоропластов *Potentilla micrantha* и *Solanum lycopersicum* (номера доступа HG931056 и HG975525 в EMBL-банке), в которых не указаны гены. В целом по рассмотренному семейству геномов на "хвост" приходится примерно десятая часть длины генома.

О ВОСЬМИКЛАСТЕРНОЙ СТРУКТУРЕ ГЕНОМОВ ХЛОРОПЛАСТОВ

Перейдем теперь к описанию структурированности, выделяемой в геномах хлоропластов различных видов. Было обнаружено, что все геномы "анфас" (первая главная компонента располагается в плоскости рисунка) имеют выраженную трёхлучевую структуру. Для значительной части геномов типичен тот вид "анфас" и "в профиль" (первая главная компонента перпендикулярна плоскости рисунка), который показан на рис. 2. Как видно (рис. 2), фрагменты из кодирующих и некодирующих участков генома образуют четко выраженные кластеры: центральную часть распределения (вид "анфас") занимают точки, соответствующие преимущественно некодирующим участкам. На одном луче лежат точки, имеющие относительные фазы F_2 и B_2, на втором луче лежат точки с относительными фазами F_0 и B_1, и на третьем лежат точки, относящиеся к относительным фазам F_1 и B_0.

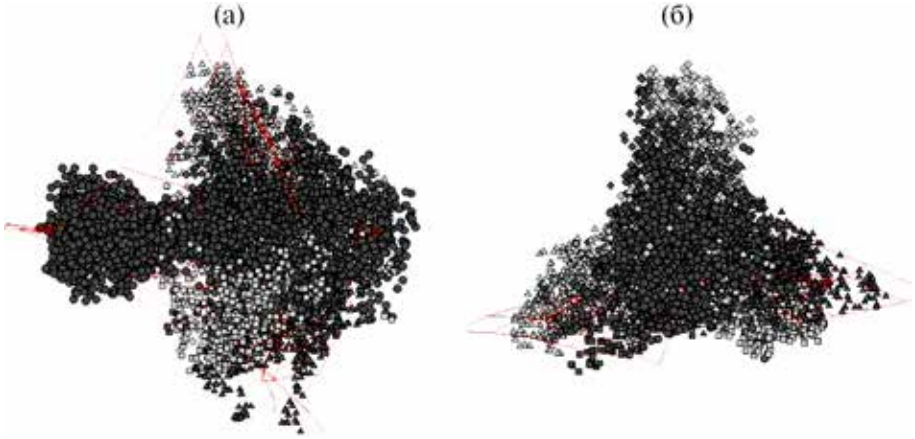

Рис. 2. Типичный вид распределения фрагментов хлоропластных геномов наземных растений по частотам троек нуклеотидов в проекциях пространства первых трех главных компонент (приведена структура генома *Lolium perenne*). Для рис. 2–5: а – вид “в профиль”, первая главная компонента лежит в плоскости рисунка; б – вид “анфас”, первая главная компонента перпендикулярна плоскости рисунка.

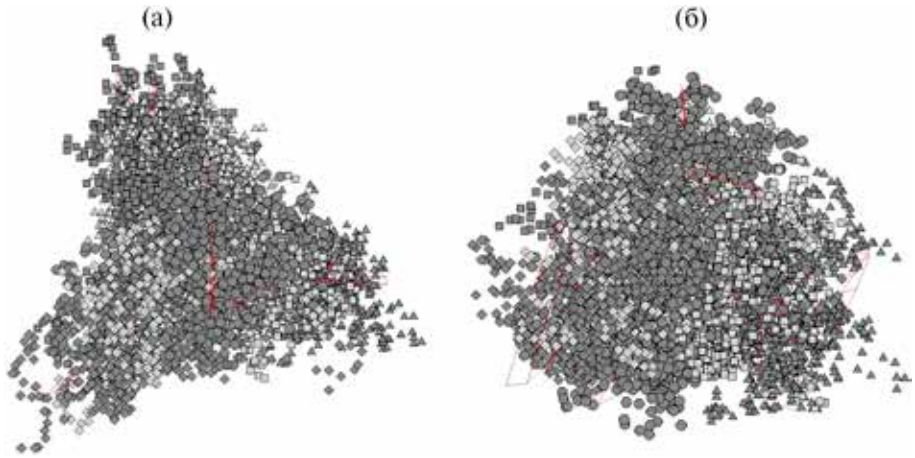

Рис. 3. Распределение фрагментов хлоропластного генома *Selaginella moellendorffii* по частотам троек нуклеотидов в проекциях пространства трех первых главных компонент.

Формально структуры, показанные на рис. 2, могут рассматриваться как четыре кластера: действительно, кластеры, соответствующие кодирующим областям из прямой и обратной цепи, весьма плотно прилегают друг к другу (в пространстве главных компонент). Мы не проводили детального исследования разделимости получаемых кластеров, однако использованными методами визуализации они отчетливо выделяются и тем самым можно говорить о существовании шести, а не трех кластеров.

На проекции “в профиль” (рис. 2, а) видно, что точки, занимающие центральную часть распределения в проекции анфас, разделены на два кластера. Первый находится в центре трёхлучевой структуры, а второй обособлен от остальных точек и именно его мы и называем “хвостом”. Такой типичной структурой данных в пространстве первых трех главных компонент обладают геномы 179 хлоропластов из 185 рассматриваемых. Собственно, наличие нового кластера – “хвоста” – позволяет утверждать существование восьмикластерной, а не семикластерной (как в работах Горбаня с соавт. (Gorban et al., 2003a, b, 2005)) структуры в геномах хлоропластов.

У оставшихся геномов хлоропластов наблюдались следующие отличия от типичной структуры данных. У геномов хлоропластов *Selaginella moellendorffii* и *S. uncinata* отсутствует обособленная группа точек, относящаяся к “хвосту”. Отсутствие хвоста у двух упомянутых видов может иметь простое объяснение: GC-состав этих геномов весьма высок и очень близок к значениям, наблюдаемым в “хвостах”. Однако такое объяснение

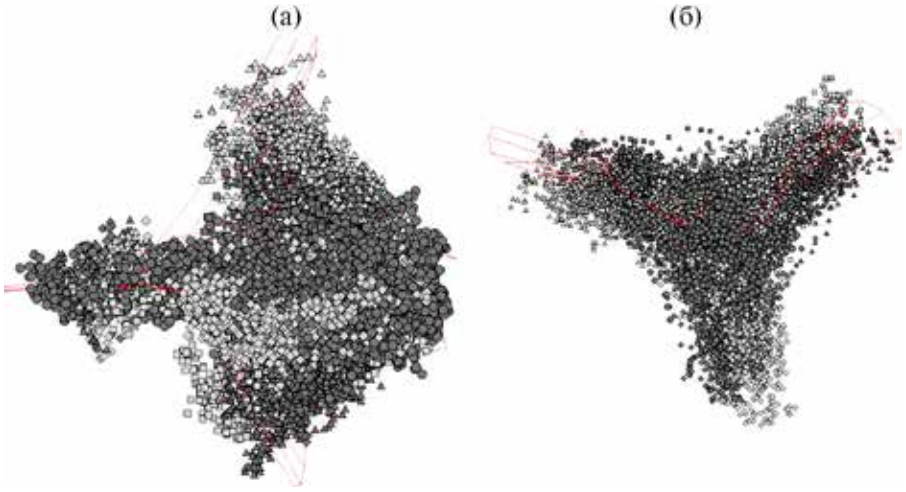

Рис. 4. Распределение фрагментов хлоропластного генома *Anthoceros* angustus по частотам троек нуклеотидов в проекциях пространства трех первых главных компонент.

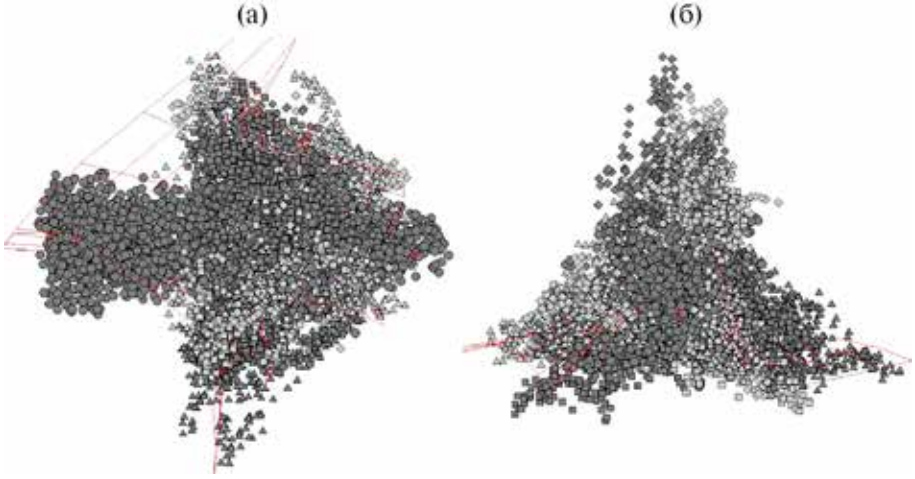

Рис. 5. Распределение фрагментов хлоропластного генома *Juniperus bermudiana* по частотам троек нуклеотидов в проекциях пространства трех первых главных компонент.

не кажется исчерпывающим, поскольку существуют организмы (например, водоросли *Aureococcus anophagefferens*, *Chromera velia* и *Cyanidioschyzon merolae*), у которых GC-контент всего генома имеет значения, близкие к 0.35, а хвост, тем не менее, отсутствует. *Selaginella moellendorffii* и *S. uncinata* – представители древних споровых растений, которые впервые появились около 400 млн лет назад. Структура этих геномов существенно больше похожа на структуру геномов бактерий (см. рис. 3). Геном хлоропласта *Anthoceros angustus* отличается от подавляющего большинства геномов хлоропластов: здесь “хвост” хлоропласта включает также фрагменты белок-кодирующих участков (рис. 4).

Это единственный представитель *Anthoceros* в рассматриваемой базе данных. У растений *Brachypodium distachyon*, *Juniperus bermudiana*, *Pinus contorta*, *Jacobaea vulgaris*, *Najas flexilis* относительные фазы в лучах перемешаны (рис. 5).

При вычислении GC-состава для геномов хлоропластов обнаружилось, что для мхов содержание GC меньше 0.3 и изменяется от 0.28361 у *Orthotrichum rogeri* до 0.2932 у *Sanionia uncinata*. Для папоротников GC-состав больше 0.4 и изменяется от 0.4064 у *Lygodium japonicum* до 0.4272 у *Myriopteris lindheimeri*. У *Selaginella moellendorffii* и *S. uncinata*, представителей древних споровых высших растений, GC-состав 0.5100 и 0.5485 соответственно. Во всех остальных геномах хлоропластов состав варьирует от 0.3 до 0.4. Таким образом, для геномов хлоропластов с GC-составом ниже 50% не установлено влияния GC-состава на

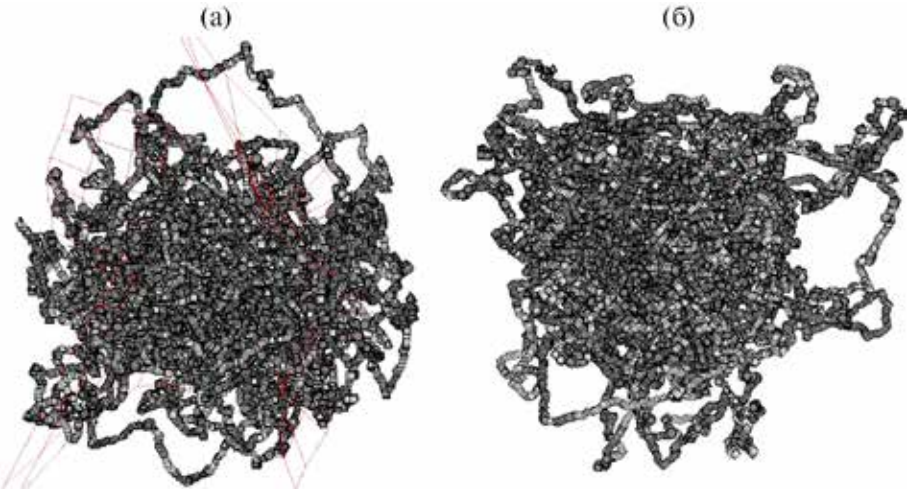

Рис. 6. Типичный вид распределение фрагментов генома цианобактерий по частотам троек нуклеотидов в проекции пространства трех первых главных компонент на примере а – *Microcystis aeruginosa NIES-843* и б – *Nostocsp. PCC7107.*

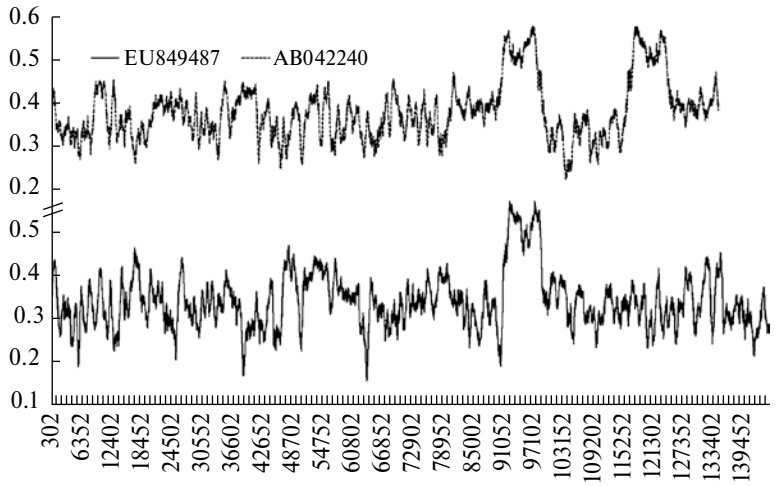

Рис. 7. Типичный график изменения GC-состава вдоль хлоропластной ДНК на примере *Triticum aestivum* (AB042240) и *Trifolium subterraneum* (EU849487).

структуру в рамках частотного словаря троек, однако этот показатель сильно коррелирует с филогенией.

Гипотеза о происхождении хлоропластов от одноклеточных свободноживущих фотосинтезирующих бактерий позволяет ожидать, что структура данных цианобактерий будет подобна аналогичной структуре хлоропластов. Мы рассмотрели структуры для геномов цианобактерий, депонированных в EMBL-банке, а именно *Microcystis aeruginosa* NIES-843, *Nostoc* sp. PCC7107, *Pleurocapsa* sp. PCC7327, *Chroococcidiopsis thermalis* PCC7203, *Gloeocapsa* sp. PCC7428, *Anabaena cylindrical* PCC7122, *Synechocystis* sp. PCC6714. Эти цианобактерии относятся к трем порядкам: Chroococcales, Nostocales, Pleurocapsales. Как видно из рис. 6, структура геномов цианобактерий существенно отличается от структуры хлоропластов. Кроме того, структура данных цианобактерий существенно отличается и от структуры иных бактерий. Семикластерная структура фрагментов 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.

Кроме кластеризации фрагментов, были построены графики изменения GC-состава вдоль хлоропластной ДНК (рис. 7). Для этого вычислялся GC-состав каждого фрагмента (напомним,

длина фрагмента $\Delta = 603$, сдвиг $t = 11$). Как видно из рис. 7, график имеет один или два пика с двумя зубцами, значения GC-состава в которых превышает 0.5. Для всех изученных последовательностей зубцы пиков совпадают с локализацией генов 16S и 23S рРНК. Иных функционально нагруженных участков в этих фрагментах не обнаружено, по крайней мере, для той аннотации геномов, которая нами использовалась. Также в терминах использованных выше обозначений, зубцы с высокой точностью соответствуют локализации фрагментов из "хвоста" генетической последовательности. Два зубца характерны для 162 последовательностей хлоропластов из 185, что соответствует расположению генов рРНК на инвертированном повторе. У 12 видов находится только один пик с двумя зубцами (*Nagei anagi*, *Trifolium subterraneum*, *Pinus contorta*, *Picea abies*, *Lathyrus sativus*, *Hordeum vulgare*, *Triticum aestivum*, *Glycyrrhiza glabra*, *Podocarpus lambertii*, *Retrophyllum piresii*, *Pseudotsuga sinensis*, *Pinus taeda*). Здесь следует подчеркнуть, что *Hordeum vulgare* и *Triticum aestivum* попали в этот список по следующим формальным причинам. В базах данных EMBL и GeneBank последовательность *Hordeum vulgare* ошибочно депонирована как полная, хотя она не содержит инвертированного повтора. Для *Triticum aestivum* в EMBL содержится две аннотации разных авторов. Одна с инвертированным повтором (AB042240), другая без него (KC912694). Оставшиеся 10 видов: *Keteleeria davidiana*, *Pelargonium* × *hortorum*, *Pisum sativum*, *Stangeria eriopus*, *Juniperus bermudiana*, *Juniperus monosperma*, *J. virginiana*, *Taxus mairei*, *Larix decidua*, *Cycas revoluta* не имеют никакой типичной картины. Для всех изученных последовательностей зубцы пиков совпадают с локализацией 16S и 23S RNA. На положение пиков на графиках влияет расположение этих rRNA в геноме. Различие графиков обусловлено различным расположением rRNA в генетических последовательностях. Иных функционально нагруженных участков в этих фрагментах не обнаружено, по крайней мере, для той аннотации геномов, которая нами использовалась. При этом генов rRNA вне хвоста не обнаружено; это верно для всех геномов, у которых хвост выделяется при кластеризации, однако хвост содержит и фрагменты, помеченные как "джанк", но при этом содержащие гены tRNA и не помеченные как "ген", поскольку длина гена tRNA заметно меньше длины фрагмента.

ОБСУЖДЕНИЕ

В работах Горбаня с соавт. (Gorban et al., 2003a, b, 2005) представлен подход к выделению структурированности бактериальных геномов, основанный на систематическом и последовательном сравнении частотных словарей фрагментов генома, выделяемых точно тем же способом, что и в нашей работе. При этом важно, что такие фрагменты также выделялись формально, вне всякой связи с их (ожидаемой) функцией либо регуляторной ролью в геноме. В этих работах показано, что частотные словари фрагментов в пространстве частот триплетов группируются в кластеры, располагающиеся в вершинах двух треугольников; вершины треугольника соответствуют сдвигу рамки считывания – иными словами, в один кластер попадают те фрагменты, которые имеют одинаковую фазу. Взаимное расположение этих треугольников, как показано в упомянутых работах, полностью определяется величиной среднего по геному GC-состава.

В указанных выше работах показано, что для бактериальных геномов наблюдается общая семикластерная структура, в которой шесть кластеров, соответствующих фрагментам, попадающим в кодирующие участки, располагаются в вершинах двух треугольников, а в их центре располагаются фрагменты, попавшие в джанк. В этих же работах приведено элегантное объяснение того, почему возникает в общем случае разбиение фрагментов на семь кластеров и при каких условиях два треугольника поворачиваются и накладываются один на другой, совпадая друг с другом и маскируясь один за другим. Ведущим фактором, определяющим величину поворота (следует отметить, что повороты там возможны не только в одной плоскости), является GC-состав генома. Причем незначительное изменение GC-состава приводит к существенному изменению взаимного расположения треугольников, соответствующих относительным фазам. При увеличении GC-состава имеют место следующие варианты структуры: "параллельные треугольники" для AT-богатых геномов (GC-состав около 25%), затем "перпендикулярные треугольники" с GC-составом около 35%, постепенно переход к вырожденному случаю в области GC = 50%, и, наконец, вырождение решения в одну плоскость, ведущую к цветкоподобному симметричному шаблону (начиная с GC = 60%). Кроме того, авторами высказывается гипотеза об универсальном характере такого рода семикластерной структуры. В работах Горбаня с соавторами (Gorban, 2003a, б, 2005) структуры геномов цианобактерий не рассматривались. Полученные нами результаты (как приведенные выше, так и не включенные в данную статью) указывают на радикальное отличие структур геномов цианобактерий от всех изученных.

Основной результат нашей работы состоит в опровержении гипотезы об универсальности семикластерной структуры любых геномов. Кроме того, для исследованных нами геномов хлоропластов наземных растений показано отсутствие влияния GC-состава генома на структуру взаимного расположения выделяемых кластеров. При значениях GC-состава от 20 до 50% структура геномов хлоропластов не менялась.

По-видимому, основная причина отличия геномов хлоропластов от геномов бактерий по тому, какую именно структуру можно наблюдать на ансамбле формально выделяемых фрагментов генома по их частотным словарям, заключается в заметно большей неоднородности геномов хлоропластов по (локальному) показателю GC-состава. Именно такие фрагменты и формируют “хвост” в распределении точек в 63-мерном пространстве частот. Следует заметить, что в геномах бактерий не существует столь обширных участков с локальным значением GC-состава, отличающихся от значения GC-состава генетической последовательности в целом. Возможно, большую неоднородность геномов хлоропластов по показателю GC-состава обуславливает то, что 16S и 23S rRNA составляют около 5% длины геномов хлоропластов, а для бактерий этот показатель существенно ниже – около 0.5%.

Структуры геномов цианобактерий вовсе не содержат кластеры. Таким образом, семикластерная структура данных, обнаруженная для бактерий, не является универсальной, что подтверждается существенным отличием аналогичных структур, наблюдаемых у хлоропластов и цианобактерий.

Авторы благодарны А.Н. Горбаню за долговременное сотрудничество в данной области. Работа выполнена при поддержке гранта № 14.Y26.31.0004 Правительства Российской Федерации. Авторы также выражают глубокую благодарность рецензенту, чьи глубокие, точные и доброжелательные замечания позволили существенно улучшить изложение результатов.

СПИСОК ЛИТЕРАТУРЫ

Мережковский К.С., 1909. Теория двух плазм как основа симбиогенезиса, нового учения о происхождении организмов. Казань: Изд-во Импер. ун-та, 102 с.

Carbonell-Caballero J., Alonso R., Ibanez V., Terol J., Talon M., Dopazo J., 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus citrus // Mol. Biol. Evol. V. 32. № 8. P. 2015–2035.

Dong W., Liu J., Yu J., Wang L., Zhou Sh., 2012. Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding // PLoS ONE. V. 7. № 4. P. 1–9.

Gielly L., Taberlet P., 1994. The use of chloroplast DNA to resolve plant phylogenies: Noncoding versus rbcL sequences // Mol. Biol. Evol. V. 11. № 5. P. 769–777.

Gorban A.N., Zinovyev A. Yu., Popova T.G., 2003a. Seven clusters in genomic triplet distributions // arXiv: cond-mat/0305681

Gorban A.N., Zinovyev A. Yu., Popova T.G., 2003b. Seven clusters in genomic triplet distributions // Silico Biol. V. 3. № 4. P. 471–482.

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

Katayama H., Ogihara Y., 1996. Phylogenetic affinities of the grasses to other monocotsas revealed by molecular analysis of chloroplast DNA // Curr. Genet. V. 29. P. 572–581.

Leliaert F., Smith D.R., Moreau H., Herron M.D., Verbruggen H., Delwiche Ch.F., De Clerck O., 2012. Phylogeny and Molecular Evolution of the Green Algae // Crit. Rev. Plant Sci. V. 31. P. 1–46.

Mereschkovsky C., 1905. Uber Natur und Ursprung der Chromatophoren im Pfl anzen reiche // Biol. Zentr.-Bl. Bd. 85. № 18. S. 593–604.

Milanowski R., Zakrys B., Kwiatowski J., 2001. Phylogenetic analysis of chloroplast small subunit rRNA genes of the genus *Euglena* Ehrenberg // Int. J. Syst. Evol. Microbiol. V. 51. P. 773–781.

Marazzi B., Endress P.K., De Queiroz L.P., Conti E., 2006. Phylogenetic relationships within *Senna* (Leguminosae, Cassiinae) based on three chloroplast dna regions: patterns in the evolution of floral symmetry and extrafloral nectaries // Am.J. Bot. V. 93. № 2. P. 288–303.

Neale D.B., Saghai-Maroof M.A., Allard R.W., Zhang Q., Jorgensen R.A., 1988. Chloroplast DNA diversity in populations of wild and cultivated barley // Genetics. V. 120. P. 1105–1110.

Shaw J., Lickey E.B., Beck J.T., Farmer S.B., Liu W., Miller J., Siripun K.C., Winder Ch.T., Schilling E.E., Small R.L., 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis // Am.J. Bot. V. 92. № 1. P. 142–166.

Eight-cluster genome structure in chloroplasts of terrestrial plants

M. G. Sadovsky[1,2], M. Yu. Senashova[1], A. V. Malyshev[1]

[1]Institute of Computational Modelling, Siberian Branch, RAS
660036 Krasnoyarsk, Academgorodok
e-mail: msad@icm.krasn.ru
[2]Siberian Federal University
School of Fundamental Biology and Biotechnology
660041 Krasnoyarsk, Svobodny Pr., 79

It is tested whether seven-cluster genome structure, detected in bacteria, is a universal one and whether all of its observed properties pertain to other genomes. Inner structuring of chloroplasts' and cyanobacteria's genome is studied, and it is found out that this structuring notably differs from what was earlier discovered in bacterial genomes. By structuring it is meant the clasterization of triplet frequency dictionary developed for different genome fragments which have regular allocation, independent of the fragment function. The clasterization has been executed by the method of elastic maps.