

УДК 573.22 + 575.89

ВЫЯВЛЕНИЕ СВЯЗИ СТРУКТУРЫ И ТАКСОНОМИИ ГЕНОМОВ ХЛОРОПЛАСТОВ МЕТОДОМ ДИНАМИЧЕСКИХ ЯДЕР

¹Садовский М.Г., ²Чернышова А.И.

¹ФГБУН «Институт вычислительного моделирования» Сибирского отделения Российской академии наук, Красноярск, e-mail: msad@icm.krasn.ru;

²ФГАОУ «Сибирский федеральный университет», Красноярск, e-mail: anna21216165@gmail.com

Представлены предварительные результаты исследования связи геномов хлоропластов с геномами организмов-носителей. 251 геном хлоропластов преобразовывался в частотные словари триплетов, затем в 63-мерном пространстве этих частот строилась классификация методом динамических ядер (до восьми классов). Полученные классы сравнивались по своему составу. Установлено, что видовой состав полученных классов не случаен: группы, полученные при делении, содержат эволюционно близкие группы геномов. Состав классов, выделяемых методом динамических ядер, оказался весьма устойчивым: доля волатильных геномов не превышала 20% в худшем случае. Эта корреляция доказывает факт сильной синхронии в эволюции геномов хлоропластов и соматических геномов, поскольку кластеризация в пространстве частот проводилась по геномам хлоропластов, а определение таксономической близости – по морфологическим признакам (т.е. по соматическому геному).

Ключевые слова: порядок, триплет, частота, классификация, корреляция, таксономия, эволюция

REVEALING THE RELATIONS BETWEEN CHLOROPLAST AND HOST GENOMES WITH K-MEANS TECHNIQUE

¹Sadovskiy M.G., ²Chernyshova A.I.

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

Siberian federal university, Krasnoyarsk, e-mail: anna21216165@gmail.com

Some preliminary results are provided approving the strong synchrony in evolution of chloroplast genomes and host genomes. To do that, 251 chloroplast genomes has been converted into frequency dictionaries of triplets, and unsupervised classification (by K-means) has been implemented. It was found the taxonomy composition of those classes is extremely regular: they form the groups of evolutionary related organisms. The stability of clusterization provided by K-means was very high: a part of volatile genomes did not exceed 20%. This correlation directly proves the strong synchrony in evolution of these two genetic systems, since the proximity in frequency space has been determined over chloroplasts, while the proximity in taxonomy has been determined morphologically (i.e. over somatic genomes).

Keywords: order, triplet, frequency, classification, correlation, taxonomy, evolution

Настоящая работа посвящена рассмотрению связи между структурой и таксономией носителей геномов на примере хлоропластов. Хлоропласты – это зеленые пластиды, по форме напоминающие двояковыпуклую линзу, как и митохондрии, способные к автономному размножению путём деления. Хлоропласт имеет собственную ДНК, которая представляет собой замкнутую кольцевую двуспиральную молекулу и рибосомы 70S-типа. Основная функция хлоропласта в организме состоит в обеспечении процессов фотосинтеза.

В настоящей работе исследуются геномы хлоропластов. Выбор таких генетических объектов обусловлен, прежде всего, тем, что у них одинаковая функция. Во-вторых, геномы хлоропластов имеют сравнительно малый размер ($\approx 10^5$ п.н.). Целью настоящей работы является установление связи между таксономией носителя хло-

ропласта и близостью двух (или более) геномов хлоропластов, определяемой исключительно по частотам встречающихся в них триплетов. Анонсируя основной результат, скажем, что такая связь выявлена; она является сильным доказательством синхронизации эволюции двух генетических систем, поскольку нуклеарный геном (определяющий таксономию организма) и геном хлоропласта непосредственно друг с другом не связаны.

Введём основные понятия. Под структурой будет пониматься частотный словарь толщины 3-символьной последовательности, соответствующей ДНК – список всех троек $v_1 v_2 v_3$ идущих подряд символов с указанием их частот; всего бывает 64 триплета. Частота – отношение числа копий n_ω данного слова к общему числу всех триплетов:

$$f_\omega = \frac{n_\omega}{N}.$$

Всякий частотный словарь отображает геном в 64-мерное метрическое пространство; близость двух геномов задается естественным образом – например, как близость двух точек в той или иной метрике. В данной работе использовалась Евклидова метрика. Один из 64 триплетов исключался, поскольку сумма всех частот в словаре равна 1. Формально исключить можно любой триплет, однако исключался тот, для которого стандартное отклонение, наблюдаемое по анализируемой выборке геномов, является минимальным. Для выявления структурной близости необходимо выделить в этом 63-мерном метрическом пространстве группы точек (геномов), которые образуют чётко выявляемые кластеры. Под таксономией понимается классическая классификация организмов по их морфологическим признакам. Следует заметить, что она определяется по соматическому (нуклеарному) геному. Тем самым связь между структурой и таксономией заключается в выявлении распределения видов, родов, семейств и т.д. по кластерам, выделяемым исключительно по частотным характеристикам рассматриваемых геномов.

Материалы и методы исследования

Данные для исследования брались на сайте www.ebi.ac.uk/genomes/organelles в базе EMBL-банка (релиз от 10 июня 2014 года). В этом релизе представлен 541 геном хлоропластов. Структура базы данных оказывает очень сильное влияние на результаты. Формальное применение методов кластеризации ко всей базе данных не приводит к построению сколько-нибудь содержательной классификации. Это связано с тем, что во всей базе данных многие таксоны высокого уровня (отряд, класс и/или выше) представлены единственным геномом (видом), а это приводит к сильнейшей «интерференции сигналов», в результате чего кластеризация не наблюдается. Для изучения связи между структурой и таксономией общую базу данных разбили на несколько отдельных подбаз. А именно, были сформированы базы частотных словарей, содержащие более 4 видов, 3 и более видов.

Кластеризацию геномов изучали методом динамических ядер. Этот метод состоит в следующем: пусть есть некоторое множество F , состоящее из подмножеств $\{M\}$. Разобьем все точки F произвольно на K классов. Для каждого класса определяем ядро (среднее арифметическое):

$$c_{v_1 v_2 v_3}^{(j)} = \frac{1}{M^{(j)}} \sum_{i=1}^{M^{(j)}} f_{v_1 v_2 v_3}(i).$$

Здесь индекс i ($1 \leq i \leq M^{(j)}$) перечисляет элементы класса; понятно, что среднее арифметическое определяется для каждого триплета $v_1 v_2 v_3$. Затем для всех точек и для каждого ядра вычислим расстояние от точки до каждого из ядер:

$$\rho^{(i)} = \sqrt{\sum_{v_1 v_2 v_3} \left(f_{v_1 v_2 v_3}(i) - c_{v_1 v_2 v_3}^{(j)} \right)^2}.$$

Индекс i теперь перечисляет все полученные классы $1 \leq i \leq K$, а индекс j перечисляет все точки множества, вне зависимости от того, к какому классу она принадлежит.

Если точка из класса **A** оказалась ближе к ядру класса **B**, то её переопределяют в этот класс. Затем снова вычислим ядра. Опять проверим все точки и так до тех пор, пока все точки не «успокоятся» – перестанут переходить из класса в класс. Они почти наверняка успокаиваются: для того, чтобы точка оказалась всё время блуждающей, нужно, чтобы точки крайне неслучайно были размещены в пространстве (см. также [1–4]).

Результаты исследования и их обсуждение

Известно, что метод динамических ядер не даёт – в общем случае – единственности построения классификации: результат зависит от начального распределения геномов по классам (которое каждый раз определяется случайным образом). Другой проблемой является определение минимального числа классов, на которое следует разделить геномы. Также не проверялась различимость классов, построенных методом динамических ядер.

Максимальное число классов, на которые разбивалось множество всех геномов, ограничивалось восемью; с одной стороны, данное ограничение носило технический характер, а с другой – кластеризация методом упругих карт [5] выделяла восемь кластеров. Тем самым естественным было остановиться на восьми кластерах как на предельном числе разбиений.

Кластеризацию методом динамических ядер можно проводить двумя способами; назовём их условно «сверху вниз» и «снизу вверх». Первый способ состоит в следующем: на первом шаге выделяем два кластера; на следующем – каждый из полученных кластеров также делим на два (либо три) кластера и так далее, до «остановки». Второй способ заключается в том, чтобы последовательно делить исходное множество на 2, 3, ..., L классов, а затем проследить судьбу геномов из j -го класса ($1 \leq j \leq R$) при переходе от разбиения на R классов к разбиению на $R - 1$ класс; здесь $L = \max\{R\}$. В рамках настоящей работы мы использовали этот второй способ («снизу вверх»).

Поскольку кластеризация методом динамических ядер зависит от начального (случайного) распределения геномов по классам, постольку результирующая кластеризация может оказаться неустойчивой: заметная доля геномов может оказаться волатильной, т.е. постоянно менять свою принадлежность к классу. В нашем случае доля волатильных геномов была невелика.

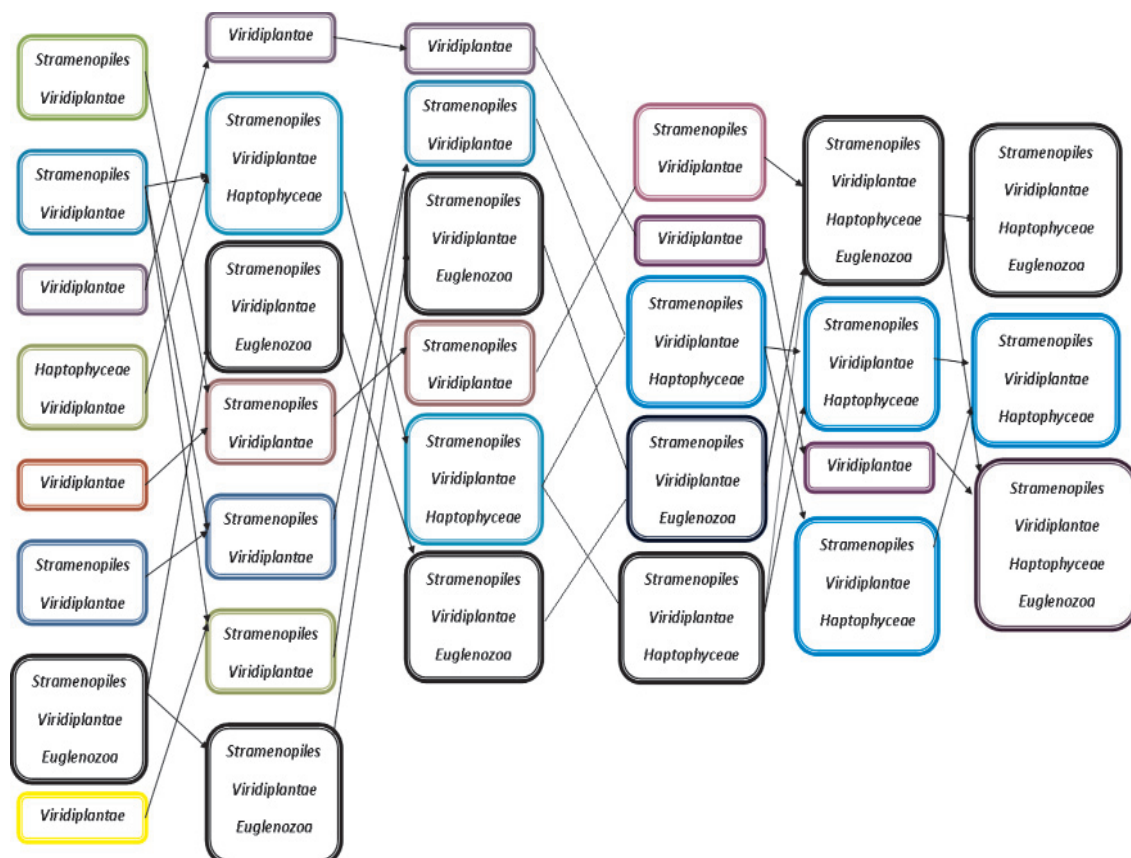


Рис. 1. Распределение таксонов высокого уровня растений, которые кластеризировались существенно неслучайно

При построении классификации «снизу вверх» возможны два диаметрально противоположных варианта: граф, соответствующий классификации, является деревом – первый вариант; во втором варианте: граф, соответствующий классификации, является полносвязным. Первый случай означает, что геномы, попадающие в какой-либо из кластеров при кластеризации на R кластеров целиком вложен в какой-то из кластеров, получающихся при кластеризации на $R - 1$ кластер. Второй случай, напротив, означает, что геномы из упомянутого кластера, полученного при кластеризации на R кластеров, (более или менее) равномерно распределяются между всеми $R - 1$ кластерами, получающимися при построении кластеризации с $R - 1$ кластером.

В нашем случае было обнаружено, что соответствующий результатам классификации граф можно представить в виде двух подграфов: первый подграф представляет собой дерево, а второй граф обладает циклами. В первый подграф, представляющий собой дерево, входили 7 таксономических групп: в первую входили *Caryophyllaceae*, *Chrysanthemum*,

Chrysobalanaceae, *Cucumis*, *Fragaria*, *Gossypium*, *Liliaceae*, *Silene*; во вторую – *Euglena*, в третью – *Nannochloropsis*, *Monodopsidaceae*; в четвертую – *Camellia*, *Corymbia*, *Cymbidium*, *Eucaliptus*, *Ginkgo*, *Magnolia*, *Nelumbo*, *Nicotiana*, *Olea*, *Picea*, *Pinus*, *Solanum*, *Vitis*; в пятую – *Hordeum*, *Oenothera*, *Oryza*, *Phyllostachys*, *Triticum*; в шестую – *Ostreococcus* и в седьмую – *Cupressaceae*, *Glycine* и *Taxus*.

На рис. 1 показана кластеризация, в результате которой образовывались устойчивые (по таксономическому составу) классы; показано распределение таксонов высокого уровня. Интересно поведение зеленых растений из третьей группы, которые на протяжении всей кластеризации не меняют свой состав, и только при разбиении на три группы к ним присоединяются другие виды: стромеопилы, гаптофитовые водоросли (гаптофиты) и евгленозои. Виды растений из седьмой группы, наоборот, распадаются, затем впоследствии снова сливаются в том же составе, и позже к ним присоединяются гаптофиты, зеленые растения, стромеопилы и евгленозои.

Список литературы

1. Горбань А. Н., Россиев Д. А. Нейронные сети на персональном компьютере. – Новосибирск: Наука, 1996. – 275 с.
2. Fukunaga K. Introduction to statistical pattern recognition. – 2nd ed. – Academic Press: London. – 591 p.
3. Горбань А.Н., Попова Т.Г., Садовский М.Г. Классификация нуклеотидных последовательностей по частотным словарям обнаруживает связь между их структурой и таксономическим положением организмов // Журнал общей биол. – 2003. – т. 64, № 5. – С. 16–21.
4. Садовский М.Г. О фундаментальной связи геномов митохондрий с геномами организмов-носителей // Фундаментальные исследования. – 2014. – № 9 (ч. 4). – С. 781–783.
5. Gorban A.N., Zinovyev A.Y. Principal manifolds and graphs in practice: from molecular biology to dynamical systems // Int. J. of Neural Systems. – 2010. – Vol. 20. – № 3. – P. 219–232.

References

1. Gorban A.N., Rossiev D.A. Neironnye seti na personal'nom komp'yutere [Neuron networks on PC]. Novosibirsk, Nauka. 1996. 275 p.
2. Fukunaga K. Introduction to statistical pattern recognition. 2nd ed. Academic Press: London. 591 p.
3. Gorban A.N., Popova T.G., Sadovsky M.G. Klassifikatsiya nukleotidnykh posledovatel'nostej po chastotnym slovarjam

obnaruzhivaet svjaz' mezhduraz' ikh strukturoj i taksonomicheskim polozheniem [Classification of nucleotide sequences over their frequency dictionaries reveals a relation between the structure of sequences and taxonomy of their bearers] // Zhurnal obshchej biologii – Journal of general biology, 2003. Vol. 64, no. 5. pp. 16–21.

4. Sadovskiy M.G. O fundamental'noj svyazi genomov mitokhondrij s genomami organizmov-nositelej [On a fundamental relation between mitochondrion and host genomes] // Fundamental'nye issledovaniya – Fundamental Researches, 2014, no. 9 (iss. 4), pp. 781–783.

5. Gorban A.N., Zinovyev A.Y. Principal manifolds and graphs in practice: from molecular biology to dynamical systems // Int. J. of Neural Systems, 2010, Vol. 20, no. 3, pp. 219–232.

Рецензенты:

Барцев С.И., д.ф.-м.н., профессор, зав. лабораторией теоретической биофизики Института биофизики СО РАН, г. Красноярск;

Кратасюк В.А., д.б.н., профессор, зав. кафедрой биофизики Института фундаментальной биологии и биотехнологии, Сибирский федеральный университет, г. Красноярск.

Работа поступила в редакцию 10.10.2014.