УДК 577.218

Рубанов Л.И., Зверков О.А., Селиверстов А.В., Любецкий В.А.

Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, г. Москва, Россия

ВЫСОКО КОНСЕРВАТИВНЫЕ ЭЛЕМЕНТЫ В МИТОХОНДРИЯХ ОДНОДОЛЬНЫХ РАСТЕНИЙ

АННОТАЦИЯ

Статья нацелена на поиск высоко консервативных элементов в геномах митохондрий однодольных растений. Полученные результаты о распределении ВКЭ хорошо согласуются с общепринятым деревом видов класса однодольных растений.

КЛЮЧЕВЫЕ СЛОВА

Большие данные; биоинформатика; высоко консервативный элемент; митохондрия; филогения; однодольные растения.

Rubanov L.I., Zverkov O.A., Seliverstov A.V., Lyubetsky V.A.

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia

HIGHLY CONSERVED ELEMENTS IN MITOCHONDRIA OF MONOCOTS

ABSTRACT

The work is aimed at the identification of highly conserved elements in mitochondrial genomes of monocotyledonous plants. The resulting distribution of identified highly conserved elements agrees well with the conventional species tree for the class Liliopsida.

KEYWORDS

Big Data; bioinformatics; highly conserved element; mitochondria; phylogeny; monocots.

Введение

Эта работа нацелена на изучение геномов митохондрий однодольных растений, класса цветковых растений, который включает многие важные сельскохозяйственные и декоративные культуры. Среди них рис, кукуруза, пшеница, сахарный тростник, лук репчатый, финиковая пальма и другие. К декоративным видам относятся, например, ирисы, тюльпаны и орхидеи. Многие виды однодольных растений служат кормом для скота. Изучение регуляции экспрессии генов митохондрий позволит понять механизмы адаптации к новым условиям окружающей среды (заморозки, засуха, засоление почв, загрязнение водоёмов).

Митохондрии — это полуавтономные органеллы, имеющие собственный геном, хотя многие белки митохондрий кодируются в ядре и приходят из цитоплазмы. Обычно в митохондриях происходит окисление жирных кислот и синтез некоторых соединений [1]. Роль митохондрий у разных организмов значительно различается, что отражается на размерах их геномов. Особенно это отличие проявляется у некоторых простейших. Например, митохондрии анаэроба *Nyctotherus ovalis* из типа Ciliophora продуцируют водород; набор белков, кодируемых в митохондриях этого вида, существенно отличается от набора у видов, живущих в аэробных условиях [2]. Обычно геномы митохондрий очень компактны, их гены образуют длинные опероны с маленькими некодирующими промежутками, поскольку механизм трансляции близок к бактериальному. Однако в митохондриях растений неожиданно большие фрагменты ДНК не кодируют ни белки, ни известные РНК. Это делает особенно интересным исследование некодирующих областей, содержащих различные элементы, регулирующие экспрессию генов.

Рассмотрим граф, рёбрам которого приписаны положительные веса. Вершины этого графа соответствуют участкам ДНК, рёбра соединяют участки с близкими последовательностями из разных геномов. «Близость» последовательностей подразумевает, что редакционное расстояние между ними не превышает заданной величины r, т.е. одну последовательность можно получить из другой последовательным применением не более чем r элементарных операций редактирования:

замены, вставки или удаления одной буквы. Если элементарные операции неравноценны, то суммарная стоимость всех операций не должна быть больше *r*.

Участки на концах любого ребра ограничиваются так, что их нельзя продолжить без того, чтобы расстояние не превысило порога *r*. При этом в качестве веса ребра используется длина этих участков (бóльшая из двух, если длины разные). Кластеры — это индуцированные подграфы, которые выбираются так, чтобы внутри кластера рёбер было больше (лучший кластер — это клика) и они имели больший суммарный вес, а между кластерами — меньше (в лучшем случае кластеры изолированные) и с меньшим весом. Каждый кластер соответствует набору достаточно длинных похожих участков сразу в нескольких геномах и называется высоко консервативным элементом (ВКЭ).

ВКЭ часто отвечают участкам генома, которые выполняют одинаковую функцию в разных организмах, причём во многих случаях эта функция неизвестна. Это объясняет интерес к нахождению ВКЭ, в том числе в сравнительно далёких друг от друга видах. Для поиска ВКЭ мы используем оригинальный метод кластеризации многодольных графов [3]. Для задачи кластеризации графа общего вида, в которой размеры кластеров ограничены сверху, недавно предложен новый приближённый алгоритм с достижимой гарантированной оценкой точности [4]. Однако этот результат относится к графам без указания весов рёбер, тогда как мы решаем более общую задачу, в которой фигурирует рёберно-взвешенный граф, что позволяет находить более осмысленные решения. Ещё один метод кластеризации MCL описан в [5], однако он требует большого времени для работы и обычно применяется для графов с малым числом рёбер, возникающих в задаче кластеризации белков.

Материалы

Геномные данные получены из базы данных GenBank. Мы рассмотрели полные геномы митохондрий 17 видов однодольных растений. Это *Allium cepa* — Лук репчатый, порядок Спаржецветные; *Phoenix dactylifera* — Финик пальчатый, порядок Пальмоцветные; *Butomus umbellatus* — Сусак зонтичный, порядок Частухоцветные; *Spirodela polyrhiza* — Многокоренник обыкновенный, порядок Частухоцветные. Прочие рассмотренные виды и подвиды относятся к порядку Злакоцветные (Poales): *Aegilops speltoides, Oryza minuta, Oryza rufipogon, Oryza sativa Indica Group, Oryza sativa Japonica Group, Sorghum bicolor, Tripsacum dactyloides, Triticum aestivum, Triticum timopheevii, Zea luxurians, Zea mays subsp. mays, Zea mays subsp. parviglumis, Zea perennis*.

Перечисленные митохондриальные геномы имеют длину в интервале 230–730 kbp, в среднем около 500 kbp. Все они состоят из одной хромосомы, которая в большинстве случаев имеет кольцевую форму (за исключением японской разновидности риса *Oryza sativa Japonica Group*, у которой эта хромосома линейная).

Методы

Дерево ВКЭ построено программой RAxML [6]. Использована модель двоичных подстановок (binary substitution model), число бутстрэп-реплик равно 300, длина ветви дерева — это среднее предсказанное число замен на позицию последовательности в ходе эволюции на данной ветви.
Поиск высоко консервативных элементов выполнен алгоритмом, который основан на поиске плотного подграфа и описан в работе [3]. Поиск потенциальных сайтов связывания транскрипционных факторов и промоторов выполнен методом, описанным в работах [7, 8]. Также использована программа MEME Suite 4.11.2 [9]. См. также http://meme-suite.org
Аннотации проверены посредством базы данных Rfam 12.1 [10].

Результаты и обсуждение

При поиске ВКЭ методом [3] мы применяли следующие значения основных параметров: длина искомых слов не менее 24, длина ключа 8, штраф за несовпадение букв 1.0, делеции не допускаются, максимально допустимое редакционное расстояние между двумя словами 3.1, число различных букв в ключе не менее 3, в слове не менее 4, число видов в кластере не менее 3.

С использованием таких параметров сначала был построен исходный граф, содержащий 218181 вершину и 290815 рёбер. После уплотнения этого графа путём объединения вершин, соответствующих сильно пересекающимся участкам, был получен начальный граф с 89468 вершинами; число рёбер не изменилось. Веса рёбер варьировались от 25 до 220, гистограмма распределения показана на рис. 1. Затем алгоритмом выделения плотных подграфов с максимальным суммарным весом рёбер было найдено 2673 кластера, каждый из которых соответствует своему ВКЭ.

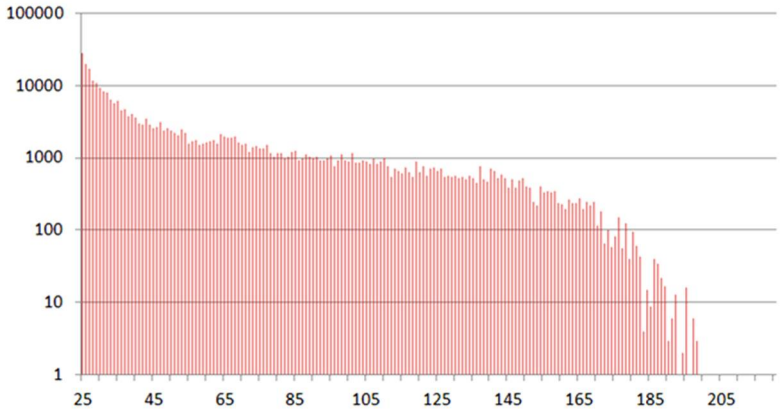

Рис. 1. Распределение рёбер начального графа по весам; по оси абсцисс указана величина веса ребра, по оси ординат – число рёбер в логарифмическом масштабе

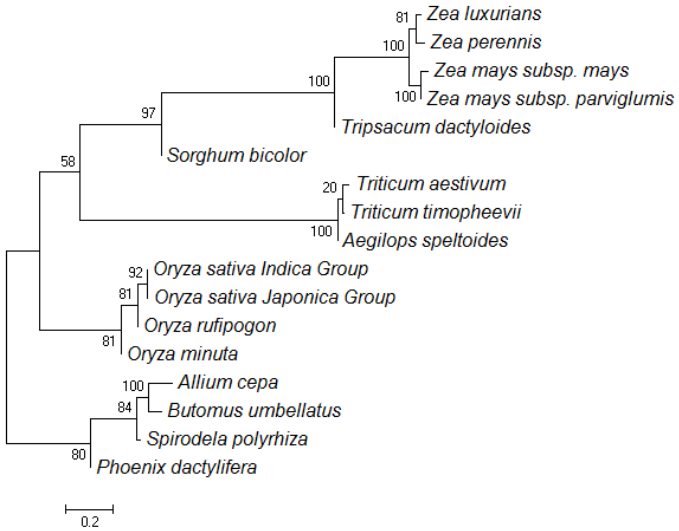

Рис. 2. Дерево высоко консервативных элементов в митохондриях однодольных растений

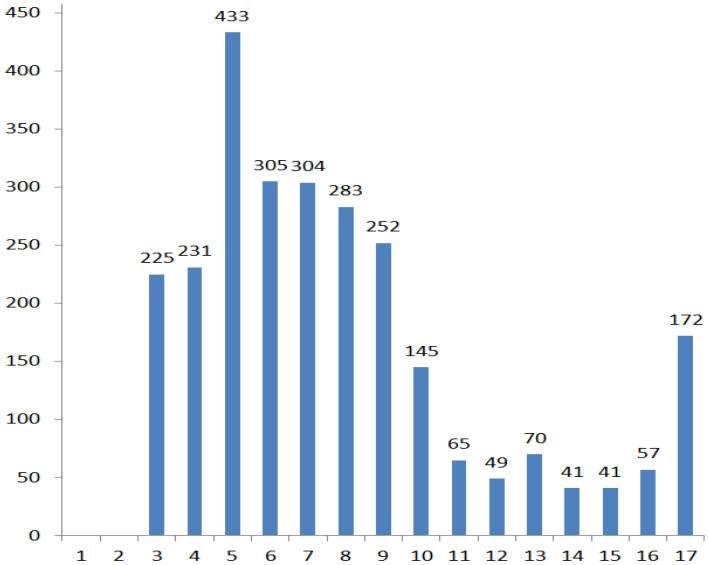

Рис. 3. Количество ВКЭ, представленных у различного числа геномов митохондрий однодольных растений; по оси абсцисс указано число видов, по оси ординат – число ВКЭ

По результатам поиска высоко консервативных элементов построено дерево (рис. 2). Это дерево построено программой RAxML на основе матрицы из 17 строк и 2673 столбцов, элемент которой 1 или 0 указывает на присутствие или соответственно отсутствие в митохондрии каждого вида представителя каждого ВКЭ. Данные о распределении ВКЭ по числу представленных в них геномов митохондрий приведены на рис. 3.

Хотя большинство митохондриальных генов выявлены давно, роль многих из них и механизмы регуляции остаются неизвестными. Дальнейшее исследование найденных ВКЭ связано с выявлением потенциальных промоторов и сайтов связывания факторов регуляции транскрипции. Для этого мы усовершенствовали ранее опубликованный алгоритм [3] путем добавления к выходным данным детальных сведений о топологии каждого кластера. Это, в частности, позволяет уточнить состав кластера, исключая слова, попавшие в результате случайных совпадений при неудачном выборе параметров. Доработка также помогла уточнить границы слов в каждой последовательности и побуквенное соответствие найденных слов. Фактически, для ВКЭ автоматически строится аналог множественного выравнивания, что повышает качество консенсусов, претендующих на роль гипотетических промоторов. Другое реализованное развитие метода — это более надёжное определение функции каждого кластера на основе всех имеющихся в GenBank аннотаций участков генома, что позволяет повысить производительность дальнейшего анализа выявленных ВКЭ. Эти и другие усовершенствования нашего метода поиска ВКЭ подробно представлены в докладе.

В рассмотренной нами задаче участвует всего 17 сравнительно коротких геномов, поэтому размерность графа получилась небольшой и для решения оказалось достаточно ресурсов обычного ПК: построение начального графа заняло около 75 мин., поиск плотных подграфов – менее минуты. Важной характеристикой нашего алгоритма кластеризации графов служит возможность эффективного использования многопроцессорных вычислительных систем – суперкомпьютеров со средой MPI [11, 12]. Это позволило нам с использованием высокопроизводительных кластеров МСЦ РАН [13] обрабатывать возникающие при работе с ядерными геномами графы значительно большей размерности, например, содержащие порядка 10^7 вершин и 10^9 рёбер [3].

В работе [14] показано, что микроэволюция пластид и митохондрий семенных растений подчиняется общим закономерностям. Полученные результаты позволяют предполагать применимость предлагаемого метода для исследования регуляции экспрессии генов не только в митохондриях, но и в пластидах растений. Рассмотренный алгоритм кластеризации графа, на котором основан поиск ВКЭ, может быть также применён для решения других задач биоинформатики и дискретной оптимизации [15-16]. В частности, большое значение имеет кластеризация белков [17–19]. В этом случае, как и при поиске ВКЭ, белки соответствуют вершинам графа, а рёбра соединяют близкие по последовательности белки.

Выводы

Полученные результаты о распределении ВКЭ хорошо согласуются с общепринятым деревом видов класса однодольных растений. Полученные результаты могут служить основой для дальнейшего предсказания регуляции экспрессии генов. Успешное применение алгоритма кластеризации графов подтверждает эффективность этого метода и открывает возможности для широкомасштабного использования при решении задач биоинформатики и дискретной оптимизации.

Литература

1. Гармаш Е.В. Митохондриальное дыхание фотосинтезирующей клетки // Физиология растений. 2016. Т. 63, № 1. С. 17–30.
2. de Graaf R.M., Ricard G., van Alen T.A., Duarte I., Dutilh B.E., Burgtorf C., Kuiper J.W., van der Staay G.W., Tielens A.G., Huynen M.A., Hackstein J.H. The organellar genome and metabolic potential of the hydrogen-producing mitochondrion of Nyctotherus ovalis // Mol. Biol. Evol. 2011. V. 28, no. 8. P. 2379–2391. DOI: 10.1093/molbev/msr059
3. Rubanov L.I., Seliverstov A.V., Zverkov O.A., Lyubetsky V.A. A method for identification of highly conserved elements and evolutionary analysis of superphylum Alveolata. BMC Bioinformatics. 2016 17:385 DOI 10.1186/s12859-016-1257-5
4. Ильев В.П., Ильева С.Д., Навроцкая А.А. О задаче кластеризации графа с ограничением на размеры кластеров // Дискретный анализ и исследование операций. 2016. Т. 23, № 3. С. 5–20. DOI: 10.17377/daio.2016.23.521
5. Van Dongen S. Graph clustering via a discrete uncoupling process // SIAM J. Matrix Anal. Appl. 2008. V. 30, № 1. P. 121–141. DOI:10.1137/040608635
6. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies // Bioinformatics. 2014. V. 30, no. 9. P. 1312–1313. DOI: 10.1093/bioinformatics/btu033
7. Селиверстов А.В., Лысенко Е.А., Любецкий В.А. Быстрая эволюция промоторов пластомных генов ndhF у цветковых растений // Физиология растений. 2009. Т. 56, № 6. С. 926–934.
8. Lyubetsky V.A., Rubanov L.I., Seliverstov A.V. Lack of conservation of bacterial type promoters in plastids of Streptophyta // Biology Direct. 2010. V. 5, no. 34. DOI: 10.1186/1745-6150-5-34
9. Bailey T.L., Bodén M., Buske F.A., Frith M., Grant C.E., Clementi L., Ren J., Li W.W., Noble W.S., MEME SUITE: tools for motif discovery and searching // Nucleic Acids Research. 2009. V. 37. P. W202-W208.
10. Nawrocki E.P., Burge S.W., Bateman A., Daub J., Eberhardt R.Y., Eddy S.R., Floden E.W., Gardner P.P., Jones T.A., Tate J., Finn R.D. Rfam 12.0: updates to the RNA families database // Nucleic Acids Research. 2015. V. 43, no. D1. P. D130–D137. DOI: 10.1093/nar/gku1063
11. Рубанов Л.И. О распараллеливании неоднородных циклов на суперкомпьютерах с распределённой памятью // Информационные процессы. 2013. Т. 13, № 4. С. 295–305.

12. Рубанов Л.И. Параллельное моделирование Монте-Карло на системах с распределённой памятью // International Journal of Open Information Technologies. 2014. Т. 2, № 2. С. 12–20.
13. Межведомственный суперкомпьютерный центр Российской академии наук. http://www.jscc.ru/scomputers.html.
14. Зверков О.А., Русин Л.Ю., Селиверстов А.В., Любецкий В.А. Изучение вставок прямых повторов в микроэволюции митохондрий и пластид растений на основе кластеризации белков // Вестник Московского университета. Серия 16. Биология. 2013. № 1. С. 8–13. DOI: 10.3103/S0096392513020119
15. Селиверстов А.В. Замечание о неявно заданных гиперграфах // Вестник Тамбовского университета. Серия: Естественные и технические науки. 2015. Т. 20, № 5. С. 1422–1424.
16. Селиверстов А.В. Многогранники и связные подграфы // Дискретный анализ и исследование операций. 2014. Т. 21, № 3. С. 82–86.
17. Sunshin Kim, KyuBum Kwack. A fast comparison algorithm to measure the accuracy of ortholog clusters // Current Bioinformatics. 2016. V. 11, no. 3. P. 324–329. DOI: 10.2174/1574893611666160322233309
18. Зверков О.А., Селиверстов А.В., Любецкий В.А. Белковые семейства, специфичные для пластомов небольших таксономических групп водорослей и простейших // Молекулярная биология. 2012. Т. 46, № 5. С. 799–809.
19. Любецкий В.А., Селиверстов А.В., Зверков О.А. Построение разделяющих паралоги семейств гомологичных белков, кодируемых в пластидах цветковых растений // Математическая биология и биоинформатика. 2013. Т. 8, № 1. С. 225–233.

References

1. Garmash E.V. Mitochondrial respiration of the photosynthesizing cell. //Russian Journal of Plant Physiology. 2016. V. 63, no. 1. P. 13–25. DOI: 10.1134/S1021443715060072
2. de Graaf R.M., Ricard G., van Alen T.A., Duarte I., Dutilh B.E., Burgtorf C., Kuiper J.W., van der Staay G.W., Tielens A.G., Huynen M.A., Hackstein J.H. The organellar genome and metabolic potential of the hydrogen-producing mitochondrion of Nyctotherus ovalis // Mol. Biol. Evol. 2011. V. 28, no. 8. P. 2379–2391. DOI: 10.1093/molbev/msr059
3. Rubanov L.I., Seliverstov A.V., Zverkov O.A., Lyubetsky V.A. A method for identification of highly conserved elements and evolutionary analysis of superphylum Alveolata. BMC Bioinformatics. 2016 17:385 DOI 10.1186/s12859-016-1257-5
4. Il'ev V.P., Il'eva S.D., Navrotskaya A.A. Graph clustering with a constraint on cluster sizes // Diskretnyi analiz i issledovanie operatsii. 2016. V. 23, no. 3. P. 5–20. DOI: 10.17377/daio.2016.23.521
5. Van Dongen S. Graph clustering via a discrete uncoupling process // SIAM J. Matrix Anal. Appl. 2008. V. 30, no. 1. P. 121–141. DOI: 10.1137/040608635
6. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies // Bioinformatics. 2014. V. 30, no. 9. P. 1312–1313. DOI: 10.1093/bioinformatics/btu033
7. Seliverstov A.V., Lysenko E.A., Lyubetsky V.A. Rapid evolution of promoters for the plastome gene ndhF in flowering plants // Russian Journal of Plant Physiology. 2009. V. 56, no. 6. P. 838–845. DOI: 10.1134/S1021443709060144
8. Lyubetsky V.A., Rubanov L.I., Seliverstov A.V. Lack of conservation of bacterial type promoters in plastids of Streptophyta // Biology Direct. 2010. V. 5, no. 34. DOI: 10.1186/1745-6150-5-34
9. Bailey T.L., Bodén M., Buske F.A., Frith M., Grant C.E., Clementi L., Ren J., Li W.W., Noble W.S., MEME SUITE: tools for motif discovery and searching // Nucleic Acids Research. 2009. V. 37. P. W202-W208.
10. Nawrocki E.P., Burge S.W., Bateman A., Daub J., Eberhardt R.Y., Eddy S.R., Floden E.W., Gardner P.P., Jones T.A., Tate J., Finn R.D. Rfam 12.0: updates to the RNA families database // Nucleic Acids Research. 2015. V. 43, no. D1. P. D130–D137. DOI: 10.1093/nar/gku1063
11. Rubanov L.I. Parallelization of nonuniform loops in supercomputers with distributed memory // Journal of Communications Technology and Electronics. 2014. V. 59, no. 6. P. 639–646.
12. Rubanov L.I. Parallel Monte Carlo modeling on distributed memory systems // International Journal of Open Information Technologies. 2014. V. 2, no. 2. P. 12–20 (in Russian).
13. Joint Supercomputer Center of the Russian Academy of Sciences. http://www.jscc.ru/scomputers.shtml.
14. Zverkov O.A., Rusin L.Yu., Seliverstov A.V., Lyubetsky V.A. Study of direct repeats in micro evolution of plant mitochondria and plastids based on protein clustering // Moscow University Biological Sciences Bulletin. 2013. V. 68, no. 2. P. 58–62. DOI: 10.3103/S0096392513020119
15. Seliverstov A.V. A note on implicitly given hypergraphs // Tambov University Reports. Series: Natural and Technical Sciences. 2015. V. 20, no. 5. P. 1422–1424 (in Russian).
16. Seliverstov A.V. Polytopes and Connected Subgraphs // Diskretnyi analiz i issledovanie operatsii. 2014. V. 21, no. 3. P. 82–86 (in Russian).
17. Sunshin Kim, KyuBum Kwack. A fast comparison algorithm to measure the accuracy of ortholog clusters // Current Bioinformatics. 2016. V. 11, no. 3. P. 324–329. DOI: 10.2174/1574893611666160322233309
18. Zverkov O.A., Seliverstov A.V., Lyubetsky V.A. Plastid-encoded protein families specific for narrow taxonomic groups of algae and protozoa // Molecular Biology. 2012. V. 46, no. 5. P. 717–726. DOI: 10.1134/S0026893312050123
19. Lyubetsky V.A., Seliverstov A.V., Zverkov O.A. Elaboration of the homologous plastid-encoded protein families that separate paralogs in Magnoliophytes // Mathematical Biology and Bioinformatics. 2013. V. 8, no. 1. P. 225–233 (in Russian).

Поступила 15.10.2016

Об авторах:

Рубанов Лев Израилевич, лаборатория № 2 Института проблем передачи информации им. А.А. Харкевича Российской академии наук, кандидат технических наук, rubanov@iitp.ru;

Зверков Олег Анатольевич, лаборатория № 6 Института проблем передачи информации им. А.А. Харкевича Российской академии наук, кандидат физико-математических наук, zverkov@iitp.ru;

Селиверстов Александр Владиславович, лаборатория № 6 Института проблем передачи информации им. А.А. Харкевича Российской академии наук, кандидат физико-математических наук;

Любецкий Василий Александрович, заведующий лабораторией № 6 Института проблем передачи информации им. А.А. Харкевича Российской академии наук, доктор физико-математических наук, lyubetsk@iitp.ru.