МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. Ломоносова
ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ

ISSN 2411-1473

Современные информационные технологии и ИТ-образование

Научный журнал

Том 2 (№ 11)

Москва
2015

УДК [004:377/378](063)
ББК 74.5(0)я431+74.6(0)я431+32.81(0)я431
С 56

Современные информационные технологии и ИТ-образование. Т. 2 (№ 11), 2015. - 614 с. (ISSN 2411-1473)

В данном выпуске журнала представлены доклады X Юбилейной международной научно-практической конференции «Современные информационные технологии и ИТ-образование», прошедшей в Московском государственном университете имени М.В. Ломоносова 20-22 ноября 2015 года.

Журнал «Современные информационные технологии и ИТ-образование» включен в наукометрическую базу «Российский индекс научного цитирования» с размещением полнотекстовых версий в научной электронной библиотеке eLIBRARY.RU. URL: http://elibrary.ru/title_about.asp?id=52785

Издание осуществлено при финансовой поддержке Российского фонда фундаментальных исследований (Грант РФФИ № 15-07-20760_г)

Учредитель:
Фонд содействия развитию интернет-медиа, ИТ-образования, человеческого потенциала «Лига интернет-медиа»

Издатель:
Фонд содействия развитию интернет-медиа, ИТ-образования, человеческого потенциала «Лига интернет-медиа»

Адрес редакции:
119991, г. Москва, ГСП-1, Ленинские горы, д. 1, стр. 52, факультет ВМК МГУ имени М.В. Ломоносова, каб. 375. E-mail: sukhomlin@mail.ru, тел./факс (495) 939-46-26.

Журнал зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор).
Свидетельство о регистрации средства массовой информации ПИ № ФС77-61433 от 10 апреля 2015 г.
Издается с 2005 года. Выходит 1 раз в год.

Рубанов Л.И.[1], Селиверстов А.В.[2], Зверков О.А.[3], Любецкий В.А.[4]

[1]ИППИ РАН, г. Москва, к.т.н., в.н.с., rubanov@iitp.ru

[2]ИППИ РАН, г. Москва, к.ф.-м.н., в.н.с., slvstv@iitp.ru

[3]ИППИ РАН, г. Москва, к.ф.-м.н., н.с., zverkov@iitp.ru

[4]ИППИ РАН, г. Москва, д.ф.-м.н., зав.лаб., lyubetsk@iitp.ru

УЛЬТРАКОНСЕРВАТИВНЫЕ ЭЛЕМЕНТЫ У ПРОСТЕЙШИХ ИЗ НАДТИПА ALVEOLATA

КЛЮЧЕВЫЕ СЛОВА

Ультраконсервативный элемент, простейшие, Alveolata, плотный подграф, кластеризация, параллельные вычисления.

АННОТАЦИЯ

Разработан и реализован алгоритм поиска ультраконсервативных элементов ДНК, основанный на поиске плотных подграфов в многодольном графе. У 22 видов из надтипа Alveolata построены кластеры ультраконсервативных элементов. Подтверждено, что род Cryptosporidium не входит в класс кокцидий, а является близким родственником плазмодиев, пироплазмид и Gregarina niphandrodes. Подтверждено, что фотосинтезирующие простейшие Chromera velia и Vitrella brassicaformis близкие родственники. Показано, что в составе типа Apicomplexa кокцидии сохранили большее число элементов, присутствовавших у общего предка надтипа Alveolata.

Введение. Ультраконсервативные элементы геномов впервые были обнаружены у млекопитающих [1, 2]. Функциональная роль таких элементов до сих пор не определена, но их очень высокая консервативность и сложность нуклеотидного состава позволяют рассматривать их как новый источник филогенетической информации в геномах, поскольку они наследуются от общего предка и находятся под действием стабилизирующего отбора. Такие участки могут служить маркёрами (зондами) для определения филогенетического положения малоизученных видов. За последнее время значительно увеличилось число секвенированных геномов простейших.

Нами проведён поиск ультраконсервативных элементов у 22 видов из надтипа Alveolata. Этот надтип представлен простейшими, многие из которых содержат пластиды, и включает тип Apicomplexa (споровики) и некоторые фотосинтезирующие виды, в том числе *Chromera velia* и *Vitrella brassicaformis*. Хотя виды из типа Apicomplexa не способны к фотосинтезу, их апикопласты происходят от пластид багрянок (Rhodophyta), в них происходит синтез изопреноидов и других соединений [3]. В то же время, криптоспоридии и близкий к ним вид *Gregarina niphandrodes* не содержат пластид [4, 5]. Подтверждение их родства позволяет, в частности, делать выводы о роли апикопластов у других споровиков. Это может иметь практическое значение, поскольку тип Apicomplexa включает большое число возбудителей протозойных инфекций, причём апикопласты служат удобной мишенью для терапевтического воздействия. Пироплазмиды представлены видами двух родов – это тейлерии и бабезии. Поскольку близкое родство тейлерий и бабезий подтверждается независимо от наших исследований, их можно рассматривать вместе. Это позволяет компенсировать погрешности, связанные с малыми размерами их геномов.

Методы. Применяемый метод состоит из двух шагов. В начале по исходным данным строится многодольный граф, доли которого соответствуют видам, вершины – участкам ДНК из этих видов, а рёбра соединяют похожие участки. При попарном выравнивании участков накладывается ограничение сверху на число идущих подряд делеций, что позволяет существенно сократить трудоёмкость вычисления. Также накладывается ограничение на суммарный штраф за замены нуклеотидов и за делеции. Кроме того участки отбирались по сложности (коэффициенту сжатия алгоритмом Лемпеля–Зива).

Пусть даны две нуклеотидные последовательности длиной n. Нужно найти все пары похожих слов длины не менее l (по одному слову из каждой последовательности). Сложность наивного алгоритма $O(n^2l^2)$ неприемлема для поиска пар слов в полных геномах, тем более для нескольких видов. Необходим быстрый алгоритм. Мы предполагаем, что слова тождественно совпадают на ключе, то есть участке длиной не менее некоторого $k < l$. Во-первых, индексируем первую последовательность, составляя хеш-таблицу всех ключей длины k в ней. Во-вторых, последовательно перебираем все ключи той же длины во второй последовательности, проверяя каждый ключ по составленной хеш-таблице. В-третьих, в случае совпадения ключей проверяем наличие в окрестности этих позиций искомой пары слов, это – ребро предварительного двудольного графа. Далее проводилась склейка вершин из одного вида, участвующих в различных ребрах, исходя из величины d перекрытия соответствующих участков генома.

Счет проводился для значений параметров: l=60, k=16, d=40. При выравнивании участков использовался штраф 1 за замену буквы и штраф 2,1 за делецию, причем допускалось не более 2 делеций подряд. Участки достаточной длины признавались похожими, если суммарный штраф за выравнивание не превышал 17,5. В точках совпадения ключей выбирались максимально длинные участки в пределах такого суммарного штрафа.

Сопоставление 22 полных геномов с построением начальных ребер графа выполнялось на суперкомпьютерах МВС-100К и МВС-10П Межведомственного суперкомпьютерного центра РАН (www.jscc.ru) и заняло около 200 часов. Последующая обработка осуществлялась на собственном 64-ядерном сервере, при этом сборка графа потребовала 13 часов на 16 процессорах, а поиск плотных подграфов тоже 13 часов, но на 22 процессорах. Отметим, что значительное время тратится на чтение и запись огромных файлов. Изначально в графе было 45 млн. вершин и 1.5 млрд. дуг (каждое ребро представлено двумя дугами встречных направлений). После объединения нескольких сильно пересекающихся участков в одну вершину их стало без малого 5 млн. (при том же числе ребер; кратные ребра удаляются алгоритмом построения плотного подграфа).

Ультраконсервативные элементы соответствуют максимальным по включению m-плотным подграфам этого графа для значений m, начиная с трёх. Напомним, что подграф многодольного графа называется m-плотным, если каждая его вершина соединена хотя бы одним ребром с вершинами из не менее чем (m–1)й другой доли. Например, m-клика в многодольном графе (с одной вершиной в доле) – это m-плотный подграф. Полный m-дольный граф m-плотный. Поиск m-клики в m-дольном графе, каждая доля которого содержит хотя бы по три вершины, – это NP-трудная задача. Однако высокая вычислительная сложность показана лишь для графов с очень высокой плотностью рёбер.

Значительная трудность состоит в том, что при неудачном выборе параметров или способа построения исходного графа в нём возникает гигантская связная компонента, объединяющая значительную часть вершин графа. Критериями успеха служат отсутствие гигантской компоненты и наличие m-плотных подграфов для значений m, больших половины от числа долей графа, то есть числа видов. Существование гигантской компоненты в исходном графе не обязательно служит препятствием к выделению m-плотных подграфов с малым числом вершин, когда параметр m достаточно большой. Однако при m=2 каждая компонента связности будет 2-плотным подграфом, который не разбивается алгоритмом на меньшие части. Значительные трудности можно ожидать и при других малых значениях m.

В случае двудольного графа, в котором никакие два ребра с общей вершиной не имеют одинакового веса, каждый выдаваемый нашим алгоритмом 2-плотный подграф имеет лишь одно ребро, то есть является 2-кликой. Отметим, что существуют 3-плотные 3-дольные графы, которые не содержат треугольников, то есть 3-клик. Пример показан на рис. 1, где вершины из одной доли лежат на горизонтальных прямых.

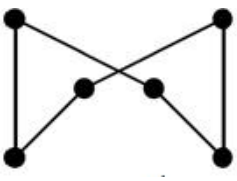

Рис.1. 3-плотный 3-дольный граф, который не содержит треугольников; вершины из одной доли лежат на горизонтали

В биоинформатике часто рассматривают взвешенные многодольные графы, у которых вес каждого ребра отражает сходство последовательностей, приписанных его концам. Важно искать

такие *m*-плотные подграфы взвешенного многодольного графа, которые содержат рёбра большого веса и не содержат много вершин из одной доли. Последние условия предполагают заданными некоторые пороги. Такие подграфы называют *кластерами*. Нами предложен алгоритм поиска кластеров, допускающий эффективное распараллеливание. Для однопроцессорного вычислительного устройства алгоритм описан в [6]. На его входе дан взвешенный многодольный граф. Этот параллельный алгоритм одновременно обрабатывает каждую вершину графа. Поскольку для алгоритма важно только взаимное соотношение между весами ребер, инцидентных одной вершине, в качестве веса каждой дуги использовалось значение длины участка, соответствующего началу этой дуги. При этом две дуги одного ребра могут иметь несколько отличающиеся веса по причине делеций, но помечаются или удаляются всегда вместе.

В каждой вершине графа алгоритм независимо выполняет следующие операции, работа начинается с шага *A*. После каждого шага выполняется синхронизация процессов во всех вершинах графа: очередной шаг начинается после окончания предыдущего шага во всех вершинах.

Шаг A. Если вершина соединена рёбрами менее чем с (*m*–1) долей, то её и все инцидентные ей рёбра удаляются. Если вершина соединена с какой-то долей лишь одним ребром, то помечаем его (в дальнейшем помеченное ребро может быть удалено только вместе с одним из его концов).

Если при выполнении шага *A* в графе произошли изменения, то снова выполняются шаги *A* во всех вершинах. В противном случае в каждой вершине однократно выполняется шаг *B*.

Шаг B. Если инцидентное вершине ребро не помечено и его вес строго меньше весов всех других не помеченных инцидентных рёбер (или оно единственное), то ребро удаляется. При проверке используются веса исходящих дуг.

Если после выполнения шага *B* в графе произошли изменения, то все вершины снова переходят к шагам *A*. В противном случае алгоритм заканчивает работу. Каждая компонента связности полученного в результате графа – искомый кластер.

Результаты. У 22 видов из надтипа Alveolata построено 845 кластеров (без гигантской компоненты). Данные о числе кластеров в зависимости от числа видов, представленных в кластере: 2 с представителями из 13 видов, 1 – из 12 видов, 4 – из 11 видов, 8 – из 10 видов, 17 – из 9 видов, 24 – из 8 видов, 35 – из 7 видов, 75 – из 6 видов, 131 – из 5 видов, 290 – из 4 видов, 258 – из 3 видов. Подавляющая доля кластеров (787) содержит не более одного участка из каждого вида, и только 9 кластеров содержат более двух участков из какого-нибудь вида (максимум 5 участков).

Большинство кластеров составляют два подмножества с маленьким пересечением: первое включает кокцидии, *Chromera velia* и *Vitrella brassicaformis*; второе включает криптоспоридии, плазмодии, пироплазмиды (тейлерии, бабезии) и *Gregarina niphandrodes*.

Вид *Gregarina niphandrodes* представлен в 10 кластерах, из которых 9 также содержат представителя из рода *Cryptosporidium*. Один из кластеров содержит два разных элемента из *Gregarina niphandrodes.*

Фотосинтезирующий вид *Vitrella brassicaformis* представлен в 34 кластерах. Из них 24 кластера одновременно содержат преставителей *Vitrella brassicaformis* и хотя бы одной кокцидии.

Другой фотосинтезирующий вид *Chromera velia* представлен в 20 кластерах, из которых 8 содержат представителей из *Vitrella brassicaformis* и 16 содержат представителей из кокцидий. То есть сходство между фотосинтезирующими видами оказалось меньше, чем сходство каждого их них с классом кокцидий в целом.

Кокцидия *Eimeria falciformis* представлена в 74 кластерах Наиболее изученная кокцидия *Toxoplasma gondii* представлена в 69 кластерах. Близкая к ней кокцидия *Neospora caninum* представлена в 53 кластерах. Кокцидия *Sarcocystis neurona* представлена в 26 кластерах, из которых 20 содержат представителей из других кокцидий и 8 содержат представителей из *Eimeria falciformis*.

Большое число кластеров (304) содержат представителей только из двух родов *Plasmodium* и *Cryptosporidium*.

Дерево, листья которого соответствуют трём видам *Gregarina niphandrodes*, *Chromera velia* и *Vitrella brassicaformis*, двум родам *Plasmodium* и *Cryptosporidium*, порядку Piroplasmida и классу Coccidia, отражающее число общих ультраконсервативных элементов, показано на рис. 2. Число в вершине дерева равно числу кластеров ультраконсервативных элементов, имеющих представителя в каждом листе порождённого поддерева. Расстояние по дереву тем меньше, чем больше общих кластеров.

Выравнивание против транскриптов из базы данных PlasmoDB (http://plasmodb.org) показывает, что наибольшие кластеры состоят из фрагментов генов, кодирующих белки. Кластеры

из 13 элементов кодируют фрагменты ribonucleoside-diphosphate reductase large subunit (PF3D7_1437200) и proline-tRNA ligase (PF3D7_1213800); кластер из 12 элементов – calmodulin (PF3D7_1434200); кластеры из 11 элементов –splicing factor U2AF small subunit (PF3D7_1119300), isoleucine-tRNA ligase (PF3D7_1332900), large subunit rRNA methyltransferase (PF3D7_1354300) и рибосомный белок L10. Здесь в скобках указан соответствующий ген из *Plasmodium falciparum* 3D7. В последнем случае фрагмент гена PF3D7_1414300 не входит в кластер, но присутствуют фрагменты ортологичных генов.

Некоторые из ультраконсервативных элементов, соответствующих вершинам гигантской компоненты графа, кодируют фрагменты рибосомных РНК или консервативных РНК сплайсосом. Однако другие не соответствуют ранее описанным типам РНК, распознаваемых сравнением с базой данных Rfam [7] или базой данных о uRNA [8]. Они требуют дальнейшего исследования. Впрочем, для сбора филогенетической информации о виде можно не выделять РНК отдельно от других участков генома. Новые участки могут представлять собой регуляторные элементы или необычные РНК.

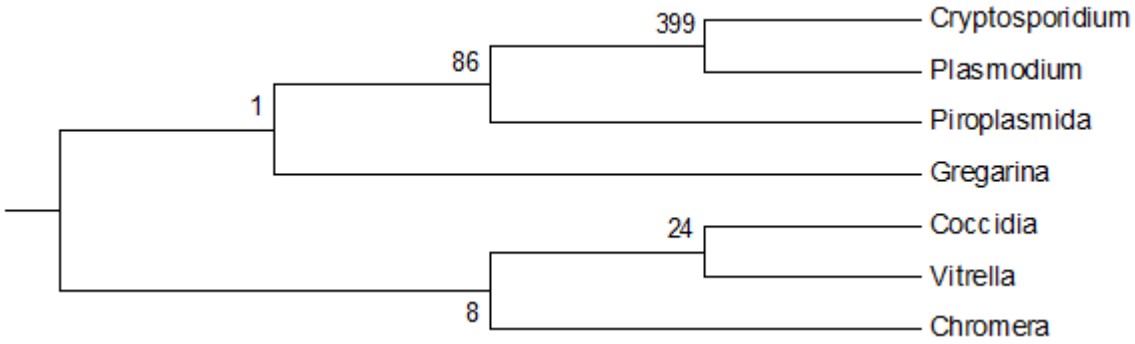

Рис.2. *Дерево видов из надтипа Alveolata, построенное по ультраконсервативным участкам. Расстояние по дереву между видами монотонно убывает при увеличении числа общих ультраконсервативных элементов* (*число общих элементов указано в корне поддерева для всех его листьев*)

Полученные результаты позволяют сделать вывод об эффективности разработанного нами алгоритма для поиска ультраконсервативных элементов генома, а также уточнить филогенетическое положение некоторых простейших из надтипа Alveolata.

Важно, что поиск ультраконсервативных участков не требует предварительной аннотации генома. Он может быть выполнен сразу после сборки достаточно длинных контигов.

Выводы

1. Подтверждено, что род *Cryptosporidium* не входит в класс кокцидий, а является близким родственником плазмодиев, пироплазмид (бабезии и тейлерии) и *Gregarina niphandrodes*. Это согласуется с недавними результатами других авторов [9, 10], хотя в прошлом криптоспоридии объединялись систематиками вместе с кокцидиями.
2. Подтверждён известный факт, что фотосинтезирующие простейшие *Chromera velia* и *Vitrella brassicaformis* близкие родственники [11].
3. В составе типа Apicomplexa кокцидии сохранили большее число элементов, присутствовавших у общего предка надтипа Alveolata. В частности, *Chromera velia* и *Vitrella brassicaformis* ближе к кокцидиям, чем к другим классам в составе типа Apicomplexa.
4. Подтверждено, что вид *Sarcocystis neurona* близок к другим кокцидиям.

Работа выполнена за счёт гранта Российского научного фонда (проект № 14–50–00150).

Литература

1. Bejerano G.., Pheasant M., Makunin I., Stephen S., Kent W.J., Mattick J.S., Haussler D. Ultraconserved elements in the human genome // *Science*, 2004, vol. 304, no. 5675, pp. 1321–13254.
2. Makunin I.V., Shloma V.V., Stephen S.J., Pheasant M., Belyakin S.N. Comparison of Ultra-Conserved Elements in Drosophilids and Vertebrates // *PloS one*, 2013, vol. 8, no. 12, e82362.
3. Садовская Т.А., Селиверстов А.В. Анализ 5'-лидерных областей некоторых генов пластид у простейших типа Apicomplexa и у красных водорослей // *Молекулярная биология*, 2009, том 43, № 4, стр. 599–604.
4. Zhu G., Marchewka M.J., Keithly J.S. *Cryptosporidium parvum* appears to lack a plastid genome // *Microbiol*. 2000, vol. 146, pp. 315–321.
5. Toso M.A., Omoto C.K. *Gregarina niphandrodes* may lack both a plastid genome and organelle // J *Eukaryot Microbiol*. 2007. vol. 54, no. 1, pp. 66–72.

6. Любецкий В.А., Селиверстов А.В. Некоторые алгоритмы, связанные с конечными группами // *Информационные процессы*, 2003, том 3, №1, стр. 39–46.
7. Burge S.W., Daub J., Eberhardt R., Tate J., Barquist L., Nawrocki E.P., Eddy S.R., Gardner P.P., Bateman A. Rfam 11.0: 10 years of RNA families // *Nucleic Acids Research*, 2013, vol. 41, Database issue, D226–232.
8. Zwieb C. The uRNA database // *Nucleic Acids Research*, 1997, vol. 25, no. 1, pp. 102–103.
9. Barta J.R., Thompson R.C.A. What is *Cryptosporidium*? Reappraising its biology and phylogenetic affinities // *Trends in Parasitology*, vol. 22, no. 10, pp. 463–468.
10. Bachvaroff T.R., Handy S.M., Place A.R., Delwiche C.F. Alveolate Phylogeny Inferred using Concatenated Ribosomal Proteins // *J. Eukaryot. Microbiol*., 2011, vol. 58, no. 3, pp. 223–233.
11. Oborník M., Lukeš J. The Organellar Genomes of *Chromera* and *Vitrella*, the Phototrophic Relatives of Apicomplexan Parasites // *Annu. Rev. Microbiol*., 2015, vol. 69, pp. 129–144.