Построение разделяющих паралоги семейств гомологичных белков, кодируемых в пластидах цветковых растений

О.А. Зверков, А.В. Селиверстов, В.А. Любецкий
ИППИ РАН
zverkov@iitp.ru

Аннотация

Разделение белков по семействам, разделяющим паралоги, позволяет уточнять аннотации белков и выполнять поиск семейства по его филогенетическому профилю, который определяется разбиением множества видов на три части. Части задают присутствие/отсутствие белка, а также случай неопределённости в этом отношении. Другое применение – поиск белков, уникальных для узкой таксономической группы («подписей»). Нами разработан алгоритм, формирующий такие семейства. Он применён к разным множествам белков. В том числе, к белкам, кодируемым в пластомах 186-ти видов цветковых растений. Полученная в этом случае база данных с возможностью поиска семейства по его филогенетическому профилю доступна по адресу http://lab6.iitp.ru/ppc/magnoliophyta/. Также алгоритм применён для разделения (кластеризации) белков, кодируемых в митохондриях 66-ти видов таксономической группы зелёных растений (Viridiplantae); соответствующая база данных: http://lab6.iitp.ru/mpc/viridiplantae/. На этой основе получены биологические результаты. Например, в митохондриях винограда (Vitis vinifera) найдены уникальные для них белки, которые в то же время типичны для пластид, что позволяет предсказать горизонтальный перенос из пластид в митохондрии.

1. Введение

Построение семейств родственных белков родофитной и хлорофитной (водоросли и мохообразные) ветвей пластид, не включающих пластиды сосудистых растений, выполнено в [1, 2]. В этой заметке рассматриваются пластиды цветковых растений и митохондрии зелёных растений.

Разделение белков на семейства (кластеризация белков) позволяет уточнять аннотации белков, выполнять поиск семейства по филогенетическому профилю, определять уникальные белки для таксономической группы; судить о работоспособности белковых комплексов, об эволюции пластомов и т.д.

Неформально говоря, задача кластеризации данного множества белков состоит в построении такого разбиения этого множества, что в один кластер попадают похожие по последовательности белки, паралоги входят в один и тот же кластер как можно реже; предполагается, что каждый кластер имеет уникального предка в дереве эволюции белков, составляющих кластер. Каждому белку заранее приписан вид, которому он принадлежит. Сложный вопрос о формальной постановке этой задачи требует дальнейших исследований; в настоящее время формальная постановка – не более чем предлагаемый алгоритм её решения.

Филогенетическим профилем называется разбиение данного множества видов на три части. Части задают присутствие/отсутствие белка (сайта связывания или другого признака вида), а также случай неопределённости в этом отношении. Вторая задача состоит в поиске кластера, который содержит только белки из видов, входящих в первую или третью части филогенетического профиля, причём для каждого вида из первой части найдётся хотя бы один белок, принадлежащий одновременно этому виду и искомому кластеру. Итак, в этой задаче дан филогенетический профиль и ищется список таких кластеров. Если задача кластеризации решена, то вторая задача решается тривиально; это относится и к задаче поиска уникальных белков. Для данного множества белков результат решения задачи кластеризации организуется в базу данных; другие задачи представляются функциями этой базы данных.

Известно несколько баз данных [3], с которыми можно сравнить так полученные наши базы данных. Однако используемые ими методы весьма трудоёмкие и, по-видимому, включают обширный «ручной» анализ, а получаемые в них кластеры включают много паралогов, значитель-

но отличающихся по последовательности. С практической точки зрения, немногие из этих баз данных включают хотя бы какие-то белки, кодируемые в пластидах, а если таковые присутствуют, то в единичных количествах. Наш алгоритм имеет квадратичную сложность от числа данных белков, и полученные семейства включают биологически мотивированные паралоги (большинство из них – точные копии друг друга).

Заметим ещё об обычных методах кластеризации: в них кластер объемлется эллипсоидом минимального объёма (эллипсоидом Левнера [4]) или сферой минимального радиуса в евклидовой метрике, или другим выпуклым множеством. Использование метрик, тем или иным образом возникающих из сходства последовательностей, приводит к различным трудностям, вызванным многозначностью геометрического образа, объемлющего кластер (как эллипсоид или сфера в упомянутых методах). Также трудности возникают из-за отсутствия выпуклости у такого образа. В нашем методе кластеризации такой образ соответствует дереву эволюции уникального предка кластера.

Наш алгоритм применён к различным наборам пластидных и митоходриальных белков, и получены соответствующие базы данных. Ниже излагается сам алгоритм и обсуждаются две новые базы данных, представляющие результаты кластеризации. Это – базы данных: 1) всех пластомных белков цветковых растений (186 полных пластомов) и 2) всех белков, кодируемых в митохондриях 66-ти видов таксономической группы зелёных растений (Viridiplantae); в обоих случаях – доступных в виде полных пластомов в GenBank, NCBI. Эти базы данных доступны по адресам http://lab6.iitp.ru/ppc/magnoliophyta/ и http://lab6.iitp.ru/mpc/viridiplantae/.

Для контроля полученных семейств (кластеров) белков использовались программа MEGA 5, [5] и база данных белковых семейств Pfam, [6].

2. Результаты: алгоритм и обоснование

Опишем оригинальный алгоритм разбиения данного множества белков на семейства. Дано множество белков (последовательностей в соответствующем алфавите), например, из пластомов родственных растений. Требуется построить кластеризацию (т.е. разбиение этого множества на попарно непересекающиеся подмножества), так чтобы в каждый кластер, максимальный по размеру, попадали сходные по последовательности белки из разных пластомов, а белки из одного пластома входили в кластер только в случае, если их сходство друг с другом больше сходства между белками из разных организмов, входящими в кластер; неформально: паралоги объединяются «как можно реже». Например, белки PsaA и PsaB, хотя имеют близкие последовательности и функционируют вместе в составе первой фотосистемы, не заменяют друг друга и нашим алгоритмом отнесены в разные кластеры. Обычные алгоритмы кластеризации не применимы к задаче кластеризации белков из-за особенностей близости, возникающей из выравнивания, а главным образом, из-за требования, относящегося к паралогам.

Наш алгоритм полезен при рассмотрении далёких видов и их белков, которые произошли от одного предкового белка и сохранили общую функцию; в этом случае сходство этих белков сравнимо или меньше сходства между паралогами. Алгоритм формирует кластеры измельчением, начиная с единственного кластера, содержащего все данные белки. Кластер может включать довольно далёкие по последовательности белки, если при этом измельчении они не попали в разные кластеры. Общий план работы алгоритма показан на рис. 1.

Пусть задан набор пластомов S_i и для каждого пластома перечислены его белки P_{ij}. Для всех пар белков (P_{ij}, P_{kl}) из всех пар пластомов вычисляется характеристика сходства $s_0(P_{ij}, P_{kl})$ белков как качество оптимального глобального выравнивания этих последовательностей; при этом само парное выравнивание не используется и не вычисляется. Эта характеристика вычисляется стандартным алгоритмом Нидлмана–Вунша [7], в котором в качестве меры сходства последовательностей, включающих делеции, используется сумма соответствующих элементов матрицы BLOSUM62, [8]. Затем алгоритм вычисляет *нормированное сходство* белков:

$$s(P_{ij}, P_{kl}) = 2s_0(P_{ij}, P_{kl})(s_0(P_{ij}, P_{ij}) + s_0(P_{kl}, P_{kl}))^{-1}.$$

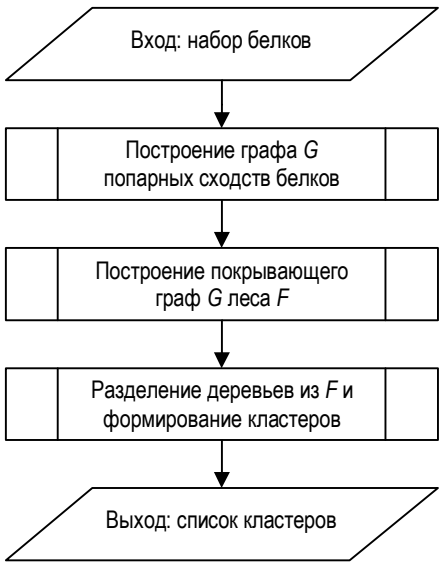

Рисунок 1. Общий план алгоритма кластеризации

Рассматривается полный неориентированный граф G_0 с множеством вершин $\{P_{ij}\}$, в котором каждому ребру (P_{ij}, P_{kl}) приписано значение $s(P_{ij}, P_{kl})$, которое будем называть весом этого ребра; рёбра соединяют различные вершины, т.е. петли отсутствуют. По G_0 строится разреженный граф G, включающий лишь рёбра (P_{ij}, P_{kl}), удовлетворяющие двум условиям: $s(P_{ij}, P_{kl}) \geq L$ и $s(P_{ij}, P_{kl}) = \max\limits_m s(P_{im}, P_{kl}) = \max\limits_m s(P_{ij}, P_{km})$, где максимумы берутся по всем белкам из соответствующих видов, *i*-го и *k*-го, а L – параметр алгоритма, по умолчанию равный нулю. В частном случае $i = k$ предполагается ещё условие $m \neq l$ и второе равенство отбрасывается.

Для полученного графа G алгоритм процедурой Крускала [9] строит лес F (ациклический подграф, компоненты связности которого – деревья), покрывающий G (рис. 2). А именно, в G перебираются рёбра в порядке убывания их веса (при совпадении весов сначала выбираются рёбра, соединяющие белки одного пластома), которые объявляются рёбрами строящегося леса F, если добавление к F очередного ребра из G не приводит к появлению в F цикла. В результате F не содержит циклов, т.е. является лесом, и включает все вершины из G. Сумма весов всех рёбер дерева называется его весом. Весом леса назовём упорядоченную по убыванию последовательность весов, составляющих его деревьев. Вес полученного леса максимален по сравнению с любым другим лесом в G.

Затем к лесу F применяется следующая процедура разделения деревьев (рис. 3), строящая набор C искомых белковых кластеров. Пусть T – дерево из F и e – ребро в T с минимальным по всем рёбрам в T весом s. Если $s < H$, где H – параметр алгоритма, и T не удовлетворяет сформулированному ниже критерию сохранения дерева, то T заменяется в F на два новых дерева T' и T'' путём удаления из T ребра e; в противном случае (т.е. критерий выполнен или $s \geq H$) дерево T перемещается из F в список C.

Критерий сохранения дерева T состоит в выполнении трёх условий (рис. 4):

(1) $|T| \leq pn$, где $|T|$ – число вершин в дереве T, n – число всех пластид в исходном наборе, p – параметр алгоритма;

(2) ребро (P_{ij}, P_{kl}) с минимальным в T весом соединяет белки P_{ij} и P_{kl}, у которых $i \neq k$;

(3) любая пара вершин (P_{ij}, P_{il}) дерева T, соответствующих белкам из *i*-й пластиды, соединена в T путём, состоящим из вершин, соответствующих белкам *i*-й пластиды (т.е. подграфы, которые состоят из вершин, относящихся к одной пластиде, связны).

Если в F ещё остались деревья, то рассматривается следующее дерево T из F, иначе алгоритм завершает работу. Полученный в результате набор деревьев C представляет собой кластеры исходных белков: один кластер состоит из последовательностей, приписанных всем вершинам одного дерева.

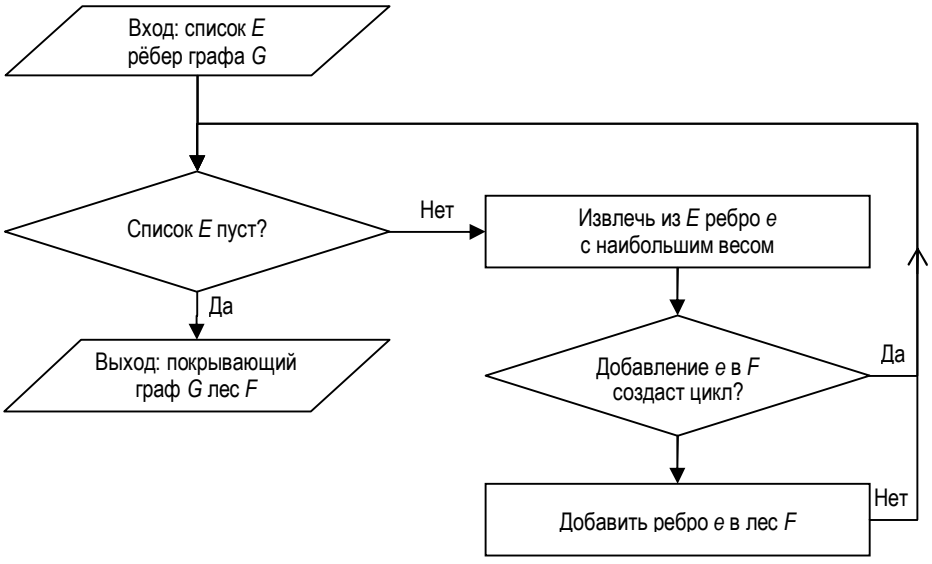

Рисунок 2. Схема алгоритма построения покрывающего леса

Вначале список E содержит все рёбра графа G, а лес F – все вершины графа G.
В результате список E пуст, а лес F накрывает все вершины графа G и его вес максимален.

Предложение 1. Пусть даны белки (последовательности в 20-буквенном алфавите) P_0 и P_n. Если среди кластеризуемых белков существует набор $\{P_i\}_{0<i<n}$, для которого при всех $i < n$ выполняется $s(P_i, P_{i+1}) > H$ и соседние белки набора соединены ребром в графе *G*, то алгоритм помещает P_0 и P_n в единый кластер.

Доказательство. Для $n = 1$ утверждение справедливо, т.к. по условию разделения алгоритм никогда не удаляет из леса рёбра с весом $s > H$. Пусть утверждение справедливо для *n*, т.е. белки P_0 и P_n принадлежат одному кластеру, и выполнено условие утверждения для $n+1$, т.е., в частности, $s(P_n, P_{n+1}) > H$. Поскольку алгоритм не удаляет из дерева рёбра с весом, $s > H$, ребро (P_n, P_{n+1}) сохраняется, т.е. белки P_n и P_{n+1} попадут в один кластер, а значит и белки P_0 и P_{n+1} попадут в один кластер. □

Предложение 2. Пусть выполнены две кластеризации C_1 и C_2 множества белков при двух значениях параметра H_1 и H_2 соответственно. Если $H_1 > H_2$, то кластеризация C_1 совпадает или является измельчением кластеризации C_2.

Доказательство. По построению кластеризации параметр *H* влияет только на принятие решения об удалении некоторых рёбер в ходе выполнения процедуры разделения, т.е., в частности, покрывающий лес, строящийся алгоритмом для данного набора белков, не зависит от *H*. При удалении каждого ребра из леса одно дерево (компонента связности, которой принадлежит удаляемое ребро) заменяется на два. Таким образом, при увеличении значения *H* каждое дерево-кластер либо останется неизменным, либо разделится на два или более кластеров, что и требовалось доказать. □

Предложение 1 описывает ограничение снизу на размер кластера. Предложение 2 неформально означает, что при увеличении параметра *H* кластеры разделяются на части, но никогда не сливаются вместе.

Следствие 1. Пусть указаны наборы белков, элементы которых предполагаются находящимися в разных кластерах. Существует не более одного числового интервала, для которого выполняется: при любом значении параметра *H* из этого интервала указанный алгоритм выдаёт набор кластеров, удовлетворяющих предположению, которые нельзя расширить (хотя бы один из них строго) с сохранением предположения. □

Следствие 2. Пусть назначен набор множеств белков, никакое множество не предполагается разделенным между разными кластерами. Существует не более одного числового интервала значений параметра *H*, для которого выполняется: при любом значении параметра *H* из этого интервала указанный алгоритм выдаёт набор кластеров, удовлетворяющих предположению, ни один из которых нельзя разбить на меньшие с сохранением предположения. □

Границы интервалов в обоих следствиях – алгоритмически вычисляемые рациональные числа. Число из пересечения этих интервалов выбирается в качестве значения параметра *H*, своего для каждой филогенетической группы. Например, у цветковых растений $H = 0.5$.

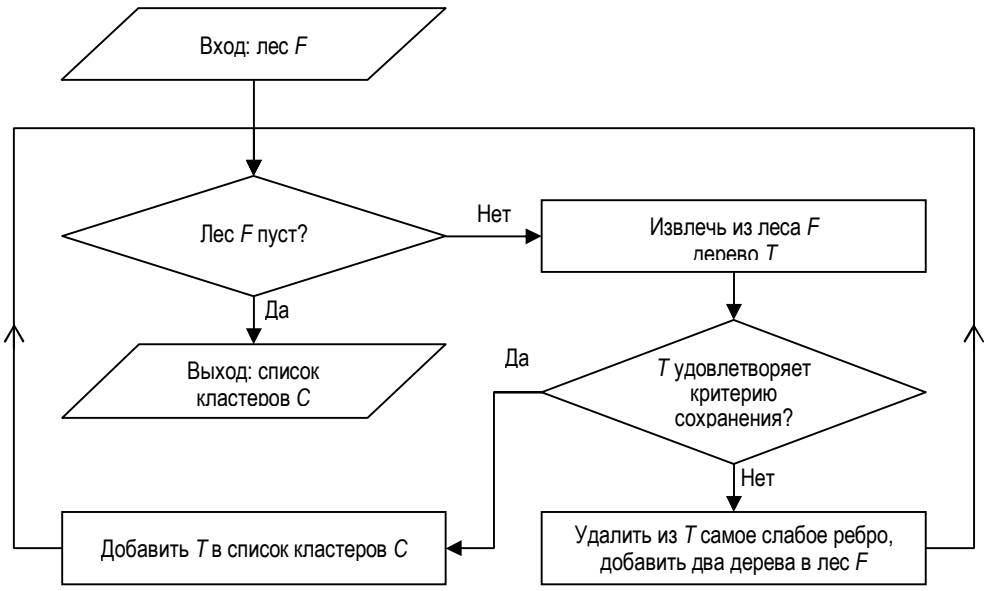

Рисунок 3. Схема алгоритма разделения леса и формирования кластеров

Вначале лес *F* содержит покрывающие *G* деревья, а список кластеров *C* пуст.
В результате лес *F* пуст, а список *C* содержит набор искомых кластеров.

3. Описание базы данных и обсуждение

В трёх случаях алгоритмически полученные кластеры были затем объединены из биологических соображений. Белок YP_003934083.1 из *Geranium palmatum* был добавлен к кластеру AccD, белок YP_654227.1 из *Oryza sativa* Indica Group – к кластеру PetE, белки YP_874745.1 из *Agrostis stolonifera* и YP_899416.1 из *Sorghum bicolor* – к кластеру Rpl23.

Веб-сайт http://lab6.iitp.ru/ppc/magnoliophyta/ обеспечивает поиск кодируемых в пластидах белков: 1) по заданному филогенетическому профилю, 2) по фрагменту аминокислотной последовательности; при этом вместе с каждой найденной последовательностью указывается кластер, которому она принадлежит.

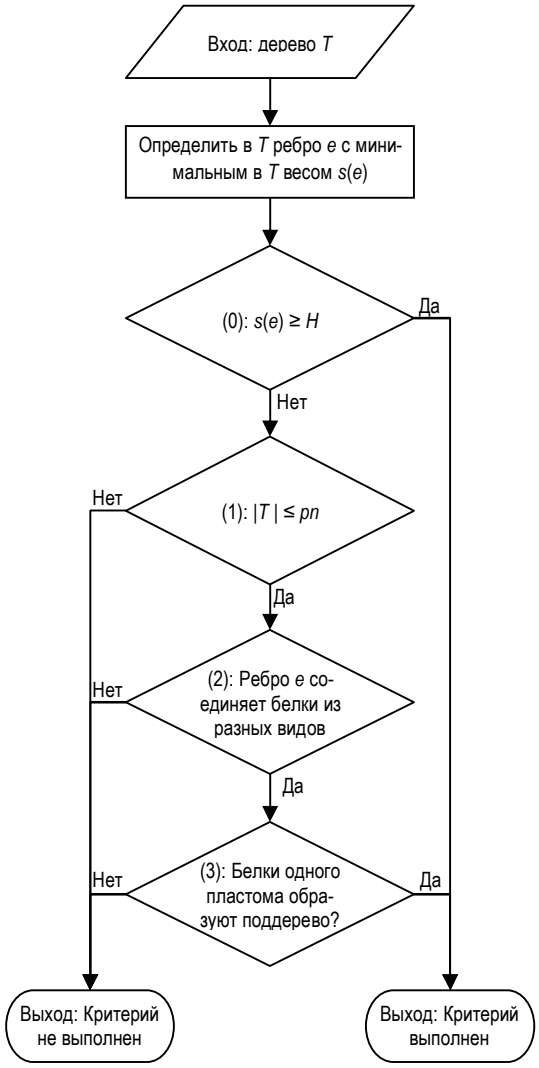

Рисунок 4. Схема проверки критерия сохранения дерева

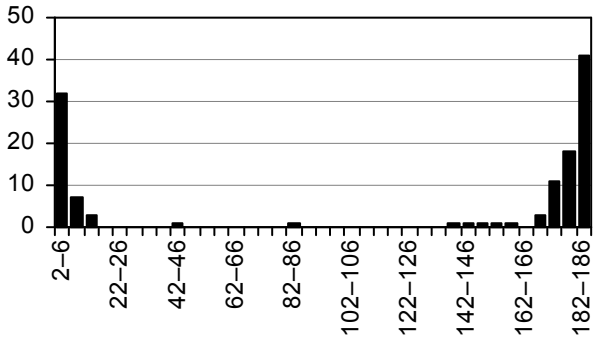

Рисунок 5. Распределение числа кластеров по их размеру

Кластеризация охватывает 15 507 белков и содержит 165 кластеров; из них 122 содержат белки из двух и более различных видов. Среди таких кластеров 39 содержат не более одного белка из каждого вида, 78 содержат пары белков из некоторых видов, но не более двух белков из каждого вида и 5 содержат более двух белков из некоторых видов, но не более четырёх белков из каждого вида.

Размер кластера понимается как число представленных в нём различных видов. Из 122-х кластеров, включающих белки из различных видов, 38 (31%) имеют размер меньше десяти, 12 (10%) имеют размер от 10-ти до 170-ти, и 72 (59%) имеют размер более 170-ти (т.е. охватывают >90% рассмотренных пластид). Чаще других встречаются кластеры с размером 182 и 183 (по 15 кластеров каждого размера). Более трети нетривиальных кластеров имеют размер больше 180-ти, т.е. каждый из них содержит белки из более чем 97% рассмотренных пластид. Распределение числа кластеров по их размеру представлено на рис. 5.

Тем же алгоритмом нами выполнена кластеризация белков, кодируемых в митохондриях 66-ти видов из таксономической группы зелёных растений (Viridiplantae). Оптимальное значение параметра для митохондрий (верхнего порога нормированного сходства белков, которые могут быть разделены между кластерами) H=0.17. Во всех данных митохондриях кодируются белки первой субъединицы цитохромоксидазы COX1, шестой субъединицы NADH-дегидрогеназы ND6 и цитохром b CytB. Многие белки уникальны для вида или нескольких близких видов. Например, у винограда (*Vitis vinifera*), наблюдаются многочисленные переносы генов, кодирующих белки фотосистем, из пластиды в митохондрию. Среди белков, типичных для пластид и кодируемых в митохондриях *Vitis vinifera*, рибосомные белки L33 (YP_002608388.1), L36 (YP_002608404.1), S15 (YP_002608362.1) и S19 (YP_002608400.1); большая субъединица рибулозо-1,5-бисфосфат карбоксилазы (YP_002608342.1); субъединица IX реакционного центра первой фотосистемы

(YP_002608389.1); белки второй фотосистемы D1 (YP_002608363.1), M (YP_002608408.1) и N (YP_002608393.1); цитохром f (YP_002608340.1); два паралога субъединицы V цитохрома b6/f (YP_002608346.1 и YP_002608390.1); субъединица VI цитохрома b6/f (YP_002608391.1); фактор InfA (YP_002608403.1); белок Ycf4 (YP_002608341.1). Такие изменения могли быть связаны с полиплоидностью ядерного генома Vitis vinifera [10]. Поиск по филогенетическому профилю белков, кодируемых в митохондриях, доступен на http://lab6.iitp.ru/mpc/viridiplantae/.

Работа выполнена при частичной финансовой поддержке Министерства образования и науки РФ (госконтракт 14.740.11.1053 и соглашение 8481).

Список литературы

[1] О.А. Зверков, А.В. Селиверстов, В.А. Любецкий, *Белковые семейства, специфичные для пластомов небольших таксономических групп водорослей и простейших*, Молекулярная биология, 2012, Т. 46, № 5, С. 799–809.

[2] О.А. Зверков, Л.Ю. Русин, А.В. Селиверстов, В.А. Любецкий *Изучение вставок прямых повторов в микроэволюции митохондрий и пластид растений на основе кластеризации белков*, Вестник Московского университета. Серия 16: Биология, 2013, № 1, С. 12–17.

[3] A.M. Altenhoff, C. Dessimoz, *Phylogenetic and functional assessment of orthologs inference projects and methods*, PLoS Comput. Biol. 5(1), 2009, e1000262.

[4] В.Л. Загускин *Об описанных и вписанных эллипсоидах экстремального объёма*, Успехи математических наук, 1958, Т. 13, № 6, С. 89–93.

[5] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods*, Molecular Biology and Evolution. 28, 2011, 2731–2739.

[6] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman, *The Pfam protein families database*, Nucleic acids research. 38, Database issue, 2010, D211-D222.

[7] Needleman S.B., Wunsch C.D. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*, Journal of Molecular Biology. 48(3), 1970, 443–453.

[8] http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt

[9] J.B. Kruskal *On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem*, Proceedings of the American Mathematical Society, 1956, V. 7. No. 1. P. 48–50.

[10] O. Jaillon, J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin *et al*. *The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla*, Nature, 2007, V. 449. No. 7161. P. 463–467.