Математическая биология и биоинформатика. 2013. Т. 8. № 1. С. 225–233.
URL: http://www.matbio.org/2013/Lyubetsky_8_225.pdf

========== МАТЕРИАЛЫ IV МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ==========
========= «МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА» ==========

УДК: 577.29

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

©2012 Любецкий В.А.[*], Селиверстов А.В.[**], Зверков О.А.[***]

Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (*ИППИ РАН*)

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 (сайта связывания или другого признака вида), 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. В этом случае соответствующая база данных и алгоритм поиска семейства по его филогенетическому профилю свободно доступны по адресу http://lab6.iitp.ru/ppc/magnoliophyta/. Алгоритм применён для разделения (кластеризации) белков, кодируемых в митохондриях 66-ти видов таксономической группы зелёных растений (Viridiplantae), и получена соответствующая база данных http://lab6.iitp.ru/mpc/viridiplantae/. Он применён к родофитной и хлорофитной (водоросли и мохообразные) ветвям пластид, и получены соответствующие базы данных, http://lab6.iitp.ru/ppc/redline/ и http://lab6.iitp.ru/ppc/chlorophyta/. 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. Формальная постановка задачи кластеризации белков, по-видимому, далека от завершения.

Ключевые слова: *кластеризация, белковые семейства, пластиды, митохондрии.*

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.

[*] lyubetsk@iitp.ru
[**] slvstv@iitp.ru
[***] zverkov@iitp.ru

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, к которому он принадлежит. Заметим, что эта постановка говорит в пользу того, что кластер не входит в минимальное объемлющее выпуклое множество. 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, которые входят 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 тривиальным алгоритмом, который мы не обсуждаем; 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 семейств белков [3]. Однако используемые авторами этих ресурсов методы весьма трудоёмки и, по-видимому, включают обширный «ручной» анализ, а получаемые в них кластеры включают много паралогов, значительно 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.

Этот алгоритм применён к различным наборам пластидных и митоходриальных белков, 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 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 программы MEGA 5 [5] и база данных белковых семейств Pfam [6].

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: последние белки входят вместе «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. Действительно, кластеры, содержащие PsaA и PsaB, имеют нормированное сходство большее, чем наименьшее нормированное сходство среди всех пар белков каждого из этих кластеров. Чтобы обеспечить расхождение PsaA и PsaB по двум разным кластерам, параметр H выбирается по Следствию 1 ниже. 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 работает быстро: за время, квадратичное от данного числа белков. Он 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.

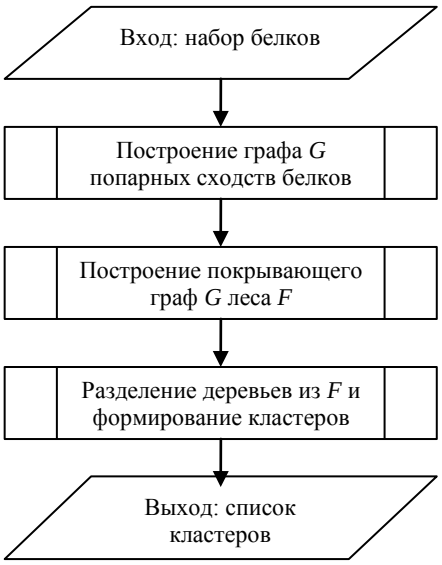

Рис. 1. Общий план алгоритма кластеризации.

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

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 графу G_0 строится разреженный граф G, включающий лишь рёбра (P_{ij}, P_{kl}), удовлетворяющие условиям: $s(P_{ij}, P_{kl}) = \max\limits_m s(P_{im}, P_{kl}) = \max\limits_m s(P_{ij}, P_{km})$ и $s(P_{ij}, P_{kl}) \geq L$, 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.

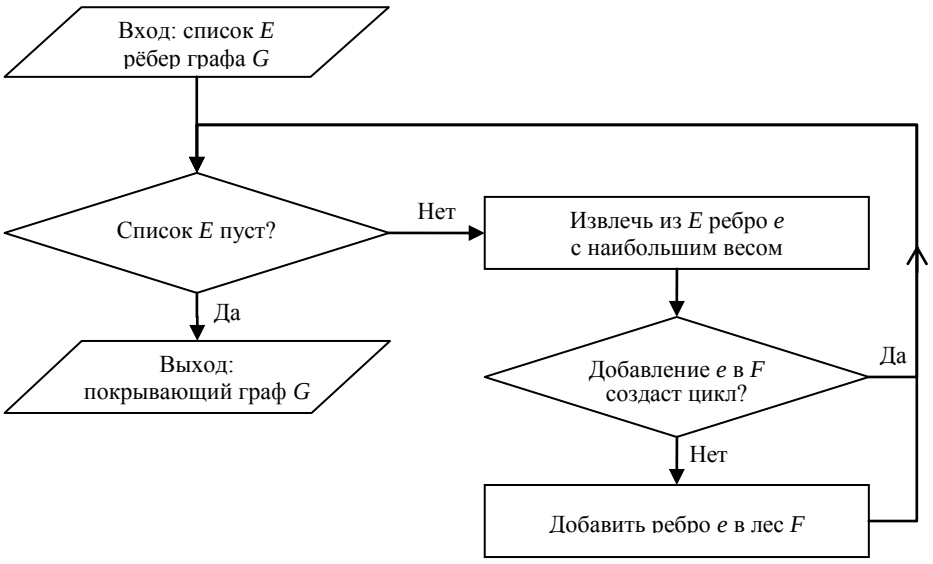

Рис. 2. Схема алгоритма построения накрывающего леса. Вначале список E содержит все рёбра графа G, а лес F – все вершины графа G. В итоге список E пуст, лес F накрывает все вершины графа G, и его вес максимален.

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.

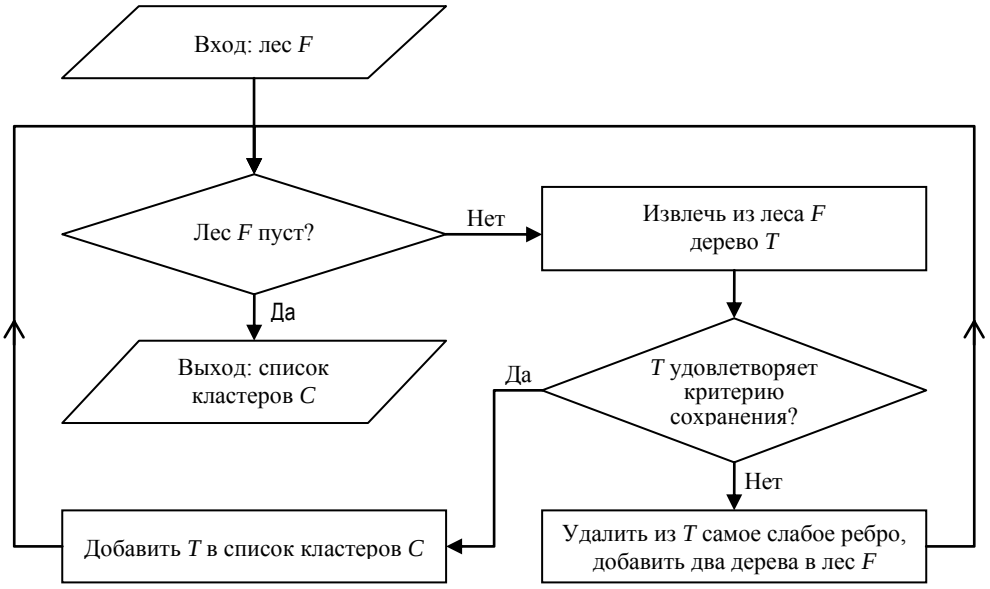

Рис. 3. Схема алгоритма разделения леса и формирования кластеров. Вначале лес F содержит накрывающие G деревья, а список кластеров C пуст. В результате лес F пуст, а список C содержит набор искомых кластеров.

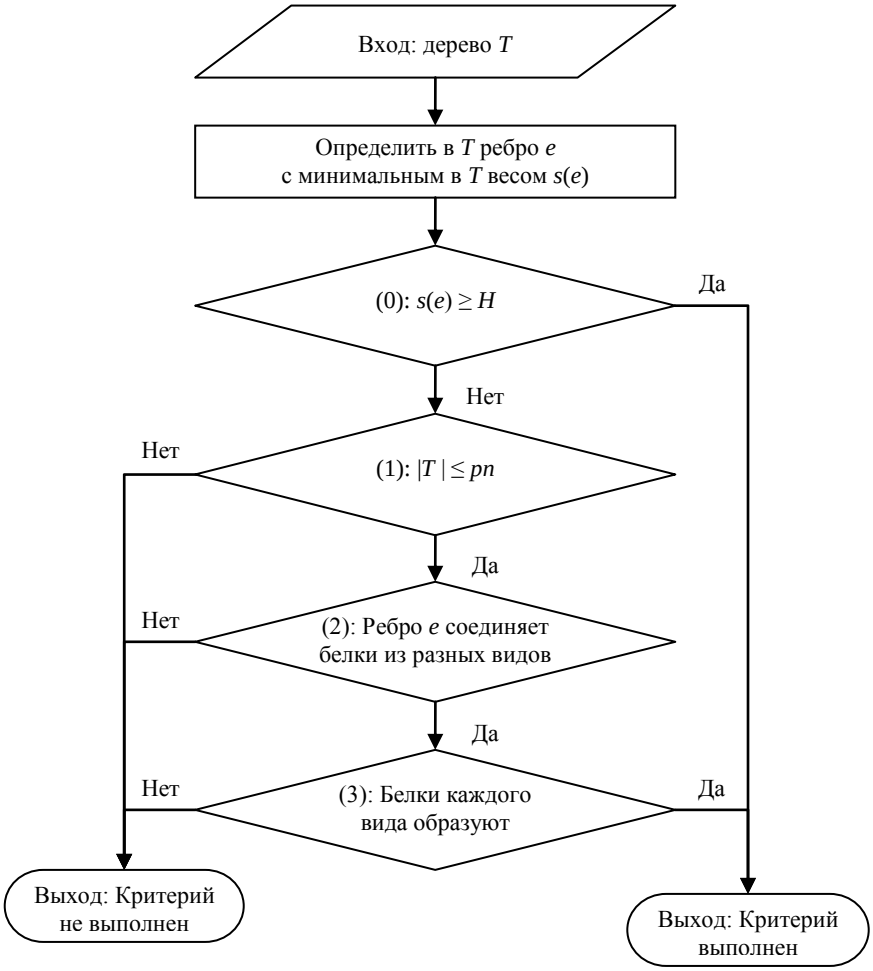

Рис. 4. Схема проверки критерия сохранения дерева.

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

1. $|T| \le pn$, 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} с $i \ne k$;
3. любая пара вершин (P_{ij}, P_{il}) дерева T, соответствующих белкам i-й пластиды, соединена в T путём, состоящим из вершин, соответствующих белкам i-й пластиды (т. е. подграфы, состоящие 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.

Предложение 1. *Пусть даны белки* P_0 *и* P_n (*последовательности в соответствующем алфавите*). *Если среди исходных белков существует набор* $\{P_i\}_{0<i<n}$, *для которого при всех* i *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 один и тот же кластер.*

Доказательство. Для $n = 1$ утверждение справедливо, так как по условию разделения алгоритм никогда не удаляет из леса рёбра с весом, превышающим 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 рёбер с весом, превышающим 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 и тот же кластер. □

Предложение 2. *Пусть выполнены две кластеризации* C_1 *и* C_2 *одного 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 процедуре разделения, т. е., в частности, покрывающий лес не зависит от H. При удалении 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. □

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 должны находиться в разных кластерах. Существует не более одного числового интервала, для которого выполняется: при любом значении параметра H из этого интервала алгоритм выдаёт набор кластеров, удовлетворяющих предположению, которые нельзя расширить* (*хотя бы один из них строго*) *с сохранением предположения.*

Следствие 2. *Пусть указаны наборы белков, для которых требуется, чтобы никакой набор не разделялся кластерами. Существует не более одного числового*

интервала, для которого выполняется: при любом значении параметра H из этого интервала 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 $H = 0.5$.

РЕЗУЛЬТАТЫ: ОПИСАНИЕ БАЗ ДАННЫХ И ОБСУЖДЕНИЕ

Всё-таки имеются три исключения, когда нам пришлось алгоритмически полученные кластеры объединить из биологических соображений. Это – белок YP_003934083.1 из *Geranium palmatum*, составлявший единичный кластер, который 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.

База данных http://lab6.iitp.ru/ppc/magnoliophyta/ обеспечивает поиск в ней кодируемых в пластидах белков: 1) по заданному филогенетическому профилю, 2) по фрагменту аминокислотной последовательности; при этом вместе с каждой из найденных последовательностей указывается кластер, к которому она принадлежит, аналогично тому, как это делает BLAST.

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

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

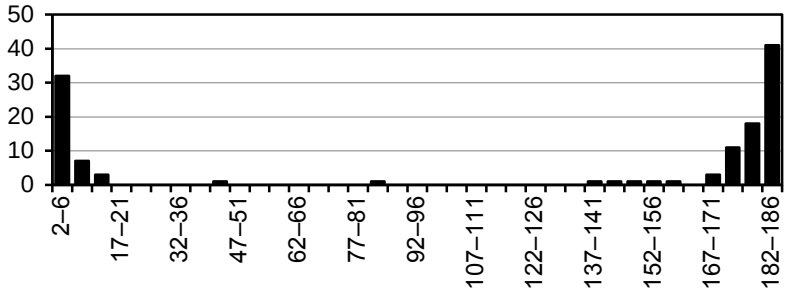

Рис. 5. Распределение числа кластеров по их размеру.

Тем же алгоритмом нами построена кластеризация белков, кодируемых в митохондриях 66 видов из таксономической группы зелёных растений (Viridiplantae) http://lab6.iitp.ru/mpc/viridiplantae/. Оптимальное значение параметра H (верхнего порога нормированного сходства белков, которым разрешено разойтись в разные кластеры) для митохондрий равно 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*. Поиск таких уникальных для вида белков – ещё одна задача, решаемая на основе нашей кластеризации. Например, у винограда (*Vitis vinifera*) наблюдаются нижеуказанные уникальные для митохондрий белки. Одновременно они интересны и тем, что указывают на многочисленные переносы генов, кодирующих белки фотосистем, из пластиды *Vitis vinifera* в её митохондрию. А именно, 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, и другие функции доступны в тех же базах данных.

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

СПИСОК ЛИТЕРАТУРЫ

1. 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.
2. 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. 12–17.
3. Altenhoff A.M., Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*. 2009. V. 5. № 1. P. e1000262. URL: http://dx.plos.org/10.1371/journal.pcbi.1000262 (дата обращения: 14.04.2013).
4. Загускин В.Л. Об описанных и вписанных эллипсоидах экстремального объёма. *Успехи математических наук*. 1958. Т. 13. № 6. С. 89–93.
5. Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*. 2011. V. 28. P. 2731–2739.
6. Finn R.D., Mistry J., Tate J., Coggill P., Heger A., Pollington J.E., Gavin O.L., Gunasekaran P., Ceric G., Forslund K. et al. The Pfam protein families database. *Nucleic acids research*. 2010. V. 38. Database issue. D211-D222. URL: http://nar.oxfordjournals.org/content/38/suppl_1/D211.full (дата обращения: 14.04.2013).
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*. 1970. V. 48. № 3. P. 443–453.
8. *BLOSUM Clustered Scoring Matrix*. URL: http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt (дата обращения: 14.04.2013).

9. Kruskal J.B. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. In: *Proceedings of the American Mathematical Society*. 1956. V. 7. № 1. P. 48–50.
10. Jaillon O., Aury J.M., Noel B., Policriti A., Clepet C., Casagrande A., Choisne N., Aubourg S., Vitulo N., Jubin C. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*. 2007. V. 449. № 7161. P. 463–467.

Материал поступил в редакцию 13.03.2012, опубликован 27.05.2013.