МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2012, том 46, № 5, с. 799–809

БИОИНФОРМАТИКА

УДК 577.21+519.163

БЕЛКОВЫЕ СЕМЕЙСТВА, СПЕЦИФИЧНЫЕ ДЛЯ ПЛАСТОМОВ НЕБОЛЬШИХ ТАКСОНОМИЧЕСКИХ ГРУПП ВОДОРОСЛЕЙ И ПРОСТЕЙШИХ

© 2012 г. О. А. Зверков*, А. В. Селиверстов, В. А. Любецкий

Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва, 127994

Поступила в редакцию 01.12.2011 г.
Принята к печати 02.03.2012 г.

Разделение белков по семействам позволяет уточнять их аннотации и искать белки по их филогенетическому профилю. Нами выполнено такое разделение (кластеризация) белков, кодируемых в пластомах багрянок и видов с пластидами, родственными пластидам багрянок (родофитная ветвь). Соответствующая база данных и поиск кластера по филогенетическому профилю белка доступны по адресу http://lab6.iitp.ru/ppc/redline. На ее основе найдены белки, специфичные для пластомов небольших таксономических групп водорослей и простейших, а также проведен поиск и анализ РНК-полимераз в ядерных геномах споровиков.

***Ключевые слова*: водоросли, споровики, пластиды, родофитная ветвь, белковые семейства, кластеры белков, филогенетический профиль.**

PROTEIN FAMILIES SPECIFIC FOR PLASTOMS IN SMALL TAXONOMY GROUPS OF ALGAE AND PROTOZOA, by *O. A. Zverkov*, A. V. Seliverstov, V. A. Lyubetsky* (Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia; *e-mail: O.Zverkov@gmail.com). Protein clustering is useful for refinement of protein annotation as well as cluster finding by its phylogenetic profile. We performed clustering of plastid encoded proteins from Rhodophyta as well as other plastid containing species related to Rhodophyta branch on species tree. Data base for cluster finding by its phylogenetic profile is available on http://lab6.iitp.ru/ppc/redline. By means of the database distinctive proteins for plastoms from small taxonomy groups of algae and protozoa were found. We performed finding and analysis of RNA polymerases encoded in Apicomplexa nuclei.

***Keywords*: algae, Apicomplexa, plastids, protein clusters, phylogenetic profile.**

Подразделение белков какой-либо таксономической группы на семейства (кластеризация белков) позволяет уточнять аннотации белков и выполнять поиск белка по его филогенетическому профилю. Такое разделение позволяет судить, например, о работоспособности белковых комплексов состоящих из нескольких субъединиц (скажем, РНК-полимераз бактериального типа). Для этого нужно проверить для данного вида наличие в нем всех необходимых субъединиц, что следует из составов кластеров каждой субъединицы. А также – позволяет судить об эволюции пластомов.

Напомним, что *филогенетическим профилем* белка называется сопоставление видам пометок: +1, если белок представлен в виде x, и –1, если он не представлен в виде x, для всех видах x из некоторого множества X; эти пометки (числа) располагаются в вектор, позиции которого помечаются индексом x; а множество X упорядочивается произвольным фиксированным образом.

Мы рассматриваем пластиды багрянок и других видов, пластиды которых родственны пластидам багрянок. Все эти виды образуют родофитную ветвь на дереве пластид [1] и так будут называться далее. Список рассмотренных пластомов приведен в табл. 1. В частности, большой интерес представляют диатомовые водоросли, пять пластомов и два полных ядерных генома которых известны. Вместе с диатомовыми мы рассмотрели два представителя надтипа Alveolata: *Durinskia baltica* (NC_014287.1) и *Kryptoperidinium foliaceum* (NC_014267.1), чьи пластомы полностью секвенированы и близки к пластому *Phaeodactylum tricornutum* [2].

Родофитная ветвь пластид включает апикопласты многих споровиков – органеллы, похожие на пластиды багрянок, но имеющие сильно

* Эл. почта: zverkov@iitp.ru

Таблица 1. Пластомы родофитной ветви

Номер локуса	Вид	Число белков	Количество кластеров	
			>1	1
NC_012898.1	*Aureococcus anophagefferens*	105	105	0
NC_012903.1	*Aureoumbra lagunensis*	110	110	0
NC_011395.1	*Babesia bovis T2Bo*	32	25	5
NC_014340.1	*Chromera velia*	80	46	31
NC_014345.1	*Chromerida sp. RM11*	81	68	6
NC_013703.1	*Cryptomonas paramecium*	82	78	4
NC_004799.1	*Cyanidioschyzon merolae strain 10D*	207	179	28
NC_001840.1	*Cyanidium caldarium*	197	185	11
NC_014287.1	*Durinskia baltica*	129	128	0
NC_013498.1	*Ectocarpus siliculosus*	148	139	5
NC_004823.1	*Eimeria tenella strain Penn State*	28	27	1
NC_007288.1	*Emiliania huxleyi*	119	117	2
NC_015403.1	*Fistulifera sp. JPCC DA0580*	135	128	4
NC_006137.1	*Gracilaria tenuistipitata var. liui*	203	193	10
NC_000926.1	*Guillardia theta*	147	143	4
NC_010772.1	*Heterosigma akashiwo*	156	138	4
NC_014267.1	*Kryptoperidinium foliaceum*	139	130	9
NC_001713.1	*Odontella sinensis*	140	132	5
NC_008588.1	*Phaeodactylum tricornutum*	132	130	0
NC_000925.1	*Porphyra purpurea*	209	208	1
NC_007932.1	*Porphyra yezoensis*	209	206	3
NC_009573.1	*Rhodomonas salina*	146	142	4
NC_014808.1	*Thalassiosira oceanica CCMP1005*	142	126	1
NC_008589.1	*Thalassiosira pseudonana*	141	127	0
NC_007758.1	*Theileria parva strain Muguga*	44	34	5
NC_001799.1	*Toxoplasma gondii RH*	26	26	0
NC_011600.1	*Vaucheria litorea*	139	139	0

Примечание. В первом столбце указан номер пластома по базе данных NCBI, во втором – вид, к которому принадлежит пластом; в третьем – число пластомных белков в этом виде, в четвертом – число кластеров пластомных белков всех видов, указанных в таблице и содержащих представителя данного вида, с числом белков строго большим 1 и соответственно равным 1.

редуцированный геном. Изучение споровиков особенно интересно, поскольку они вызывают заболевания человека и животных. В частности, *Theileria* и *Babesia* переносятся иксодовыми клещами [3] и вызывают заболевания крупного рогатого скота: *B. bigemina* и *B. bovis* – бабезиоз крупного рогатого скота, *Th. annulata* – тейлериоз крупного рогатого скота, *Th. parva* – лихорадку Восточного Берега; *Eimeria tenella* – эймериоз кур; *Toxoplasma gondii* – токсоплазмоз кошек и человека. Различные виды рода *Plasmodium* вызывают малярию у людей (*Pl. falciparum*), грызунов и других животных. Геномы *B. bovis* и *Th. parva* чрезвычайно близки между собой [4]. Особенности и функции апикопластов рассмотрены в обзоре [5]. Отметим, что некоторые споровики, например, *Cryptosporidium parvum* [6], не имеют пластид.

Исследование разнообразных процессов, связанных с апикопластами, позволит понять их роль в передаче инфекции и механизм действия лекарственных средств на апикопласт. Поскольку в апикопласте трансляция и обычно также транскрипция имеют бактериальную природу, именно апикопласты служат главной мишенью антибиотиков, не оказывающих прямого воздействия на экспрессию ядерных и митохондриальных генов, поэтому значение проблемы исследования механизмов регуляции и эволюции этих процессов велико. Некоторые результаты на эту тему содержатся в наших предыдущих работах [7, 8].

Поскольку многие белки, достигающие пластид, кодируются в ядре, необходимо сопоставлять данные о белках, образующихся в ядре, с данными о генах и регуляторных областях в пластоме. Особую роль играют субъединицы РНК-полимераз бактериального типа и РНК-полимеразы фагового типа, гомологичные РНК-полимеразам бактериофага T7 [9, 10], которые обеспечивают транскрипцию в пластидах и митохондриях [11].

Самостоятельную задачу представляет собой проблема кластеризации белков, что позволяет уточнять аннотации белков, эффективно проводить поиск белка по его филогенетическому профилю и определять возможности организма для адаптации в различных условиях. В частности, основанная на кластеризации база данных позволит решать перечисленные выше вопросы. Известно несколько таких баз [12], однако большинство из них содержит небольшое число видов, из которых лишь немногие имеют пластиды из родофитной ветви. Например, база OrthoMCL [13] (последняя версия от 31 марта 2011) включает 150 геномов, из них лишь некоторые принадлежат к рассмотренной нами родофитной ветви; база RoundUp [14] содержит небольшое число видов, среди которых всего несколько водорослей и споровиков; база OMA [15] (версия от 18 мая 2011) охватывает 1109 видов, но почти все они не содержат пластид из родофитной ветви; база EggNOG [16] в современной версии включает 1133 вида, но виды из родофитной ветви в ней не представлены; база InParanoid [17] (версия 7.0) содержит всего 100 организмов эукариот, среди которых к родофитной ветви принадлежит лишь одна диатомовая водоросль и несколько споровиков; в базах COG и KOG [18] представлено небольшое число видов, только два растения и ни одного вида из рассмотренной нами родофитной ветви.

Кластеризация белков, кодируемых в пластидах, приводит к новой базе данных, в частности, удобной для исследования споровиков – возбудителей многих протозойных инфекций.

Мы осуществили разделение (кластеризацию) белков, кодируемых в пластомах родофитной ветви. Часть соответствующей базы данных приведена в Приложении (см. дополнительные материалы на сайте www.molecbio.com/downloads/2012/5/supp_zverkov_rus.pdf). Поиск кластера по филогенетическому профилю белка в этой базе доступен на сайте [19]. На ее основе найдены белки, специфичные для пластомов небольших таксономических групп водорослей и простейших, а также РНК-полимеразы в ядерных геномах споровиков и, в частности, σ-субъединицы РНК-полимераз бактериального типа и РНК-полимераз фагового типа у видов надтипа Alveolata.

АЛГОРИТМ ВЫДЕЛЕНИЯ БЕЛКОВЫХ СЕМЕЙСТВ (КЛАСТЕРИЗАЦИИ)

Предлагаем следующий алгоритм кластеризации аминокислотных последовательностей белков (слов различной длины в 20-буквенном алфавите) по их сходству. Алгоритм, в частности, применен к исследованию пластомов водорослей и споровиков и при создании базы данных белковых семейств, представленных в пластидах.

Кластеры формируются путем «измельчения», начиная с единственного кластера, содержащего все белки. Кластер может включать далекие белки, если при измельчении они не попали в разные кластеры. Такой подход полезен при рассмотрении далеких видов и их белков, которые произошли от одного предкового белка и сохранили общую функцию, если сходство этих белков сравнимо или меньше сходства между многими паралогами (близкими гомологами из одного генома). Кроме того, наш алгоритм работает очень быстро, за время, квадратичное по отношению к общему числу видов n. Рассмотрим алгоритм.

Пусть задан набор пластид, которые индексируем буквой i, и для каждой пластиды заданы ее белки P_{ij}, индексируемые буквой j. Для всех пар белков (P_{ij}, P_{kl}) из всех пар видов (S_i, S_k) вычисляется характеристика близости $s_0(P_{ij}, P_{kl})$ белков как качество оптимального глобального выравнивания этих последовательностей; при этом само парное выравнивание не используется и может не вычисляться. Эта характеристика вычисляется с использованием стандартного алгоритма Нидлмана–Вунша [20], в котором в качестве меры сходства последовательностей используется сумма соответствующих элементов матрицы BLOSUM62 [21]. Затем алгоритм вычисляет нормированную *степень сходства* $s(P_{ij}, P_{kl})$ белков по формуле $2s_0(P_{ij}, P_{kl})(s_0(P_{ij}, P_{ij}) + s_0(P_{kl}, P_{kl}))^{-1}$. 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; ребра соединяют различные вершины, т.е. петли отсутствуют.

Для уменьшения объема вычислений вместо полного графа можно использовать *разреженный* граф G, состоящий только из ребер (P_{ij}, P_{kl}), удовлетворяющих условиям:

$$s(P_{ij}, P_{kl}) = \max_m s(P_{im}, P_{kl}) = \\ = \max_m s(P_{ij}, P_{km}) \quad \text{и} \quad s(P_{ij}, P_{kl}) \geq L, \tag{1}$$

где максимумы берутся по всем белкам из соответствующих видов i-го и k-го, а L –параметр алгоритма, по умолчанию равный 0. Если $i = k$, то

предполагается еще условие: $m \neq l$, и второе равенство нужно опустить.

В полученном графе G алгоритм строит набор всех его связных компонент. Затем для каждой компоненты (обозначим ее той же буквой G) строятся «накрывающие» G бескорневые деревья D. Последнее означает следующее: в G перебираются ребра в порядке убывания их веса, которые объявляются ребрами строящегося дерева D; если добавление к D очередного ребра из G приводит к появлению в D цикла, то это ребро пропускается. В результате D не содержит циклов, т.е. является деревом, и включает все вершины из G. Сумма весов всех ребер в D называется *весом* дерева D; полученные деревья имеют максимально возможный вес. Итак, для каждой связной компоненты в исходном графе G строятся накрывающие ее деревья D.

Затем к каждому дереву D применяется следующая рекурсивная *процедура* "разделения дерева", которая строит набор деревьев $\{D_{i,j,..}\}$, в котором все индексы равны 1 или 2. Длина последовательности i, j,… индексов называется *глубиной* дерева $D_{i,j,..}$. Если текущее дерево, скажем, $D_{i,j,..}$ некоторой глубины k из этого набора не удовлетворяет сформулированному ниже критерию разделения дерева, то оно заменяется в наборе на два дерева $D_{i,j,...,1}$ и $D_{i,j,...,2}$, глубины $k + 1$ каждое, путем удаления из $D_{i,j,..}$ ребра e_0 с минимальным по всему $D_{i,j,..}$ весом s, если выполнено условие $s < H$, где H – параметр алгоритма (если это неравенство не выполняется, то текущее дерево не делят и переходят к следующему дереву). Иначе проверяется критерий сохранения текущего дерева $D_{i,j,..}$ без изменения. Этот критерий для дерева $D_{i,j,..}$ с множеством вершин V состоит в выполнении трех условий: 1) $|И| \leq pn$, где $|И|$ – число вершин в дереве $D_{i,j,..}$, а n – число всех видов в исходном наборе видов и p – параметр алгоритма; 2) ребро (P_{mq}, P_{kl}) с минимальным весом в $D_{i,j,..}$ соединяет белки P_{mq} и P_{kl}, у которых индексы $m \neq k$; 3) любая пара вершин P_{mq} и P_{ml} дерева $D_{i,j,..}$, соответствующих белкам из одного вида, соединена в $D_{i,j,..}$ путем, состоящим из вершин, соответствующих белкам того же вида. Если этот критерий выполнен и имеется следующее дерево, то переходим к нему. Если все деревья уже исчерпаны, то алгоритм завершает работу.

Полученный в результате набор деревьев представляет собой разбиение исходных белков на кластеры, состоящие из последовательностей, приписанных всем вершинам одного дерева.

Искусственный пример, иллюстрирующий работу алгоритма

Исходные данные. Кластеризуются произвольно выбранные белки, кодируемые в трех пластомах: NC_000925 (*Porphyra purpurea*), NC_000926 (*Guillardia theta*), NC_000927 (*Nephroselmis olivacea*). А именно, из каждого пластома взято по три коротких белка:

ref|NP_053804.1| photosystem_I subunit IX [*Porphyra purpurea*]
MNNNFTKYLSTAPVIGVLWMTFTAGFIIELNRFFPDVLYFYL;
ref|NP_054005.1| photosystem_I subunit XII [*Porphyra purpurea*]
MIDDSQIFVALLFALVSAVLAIRLGKELYQ;
ref|NP_053866.1| ribosomal protein S18 [*Porphyra purpurea*]
MAVYRKKISPIKPTEAVDYKDIDLLRKFITEQGKILPKRSTGLTSKQQKKLTKAIKQARILSLLPFLNKD;
ref|NP_050719.1| photosystem_I subunit VIII [*Guillardia theta*]
MTAAYLPSILVPIIGIIFPGLTMAFAFIYIEQDQIN;
ref|NP_050713.1| photosystem_I subunit IX [*Guillardia theta*]
MDNNFLKYLSTAPVLLTIWLSFTAALVIEANRFYPDMLYFPI;
ref|NP_050701.1| photosystem_I subunit XII [*Guillardia theta*]
MISDTQIFVALILALFSFVLAIRLGTSLY;
ref|NP_050833.1| photosystem_I subunit VIII [*Nephroselmis olivacea*]
MVTSFLPSLFVPLVGLVFPAVAMASLFLYIEKDEIA;
ref|NP_050847.1| photosystem_I subunit IX [*Nephroselmis olivacea*]
MKDFTTYLSTAPVLAAVWFGFLAGLLIEINRFFPDALSFSFV;
ref|NP_050819.1| ribosomal protein L36 [*Nephroselmis olivacea*]
MKVRPSVRKICDKCCLIRRHRKLLVICSNPKHKQRQG.

Обозначим эти белки в указанном порядке как 1:1, 1:2, 1:3; 2:1, 2:2, 2:3; 3:1, 3:2, 3:3. Таким образом, пара n:m обозначает m-й белок из n-го пластома. Степени сходства s_0 всех пар белков приведены в табл. 2, нижняя треугольная часть этой таблицы совпадает с верхней, ее диагональ не используется. Нормированные степени сходства s всех пар белков приведены в табл. 3. Числовые значения округляются до двух знаков.

Положим порог L равным нулю. После отбрасывания ребер в графе G по второму условию в формуле (1) остается 15 чисел в табл. 4. После отбрасывания ребер в графе G по первому условию в формуле (1) остаются 8 значений, отмеченных в

Таблица 2. Степени сходства s_0 пар белков

s_0	1:1	1:2	1:3	2:1	2:2	2:3	3:1	3:2	3:3
1:1	225	−9	−60	11	153	1	7	131	−36
1:2	−9	139	−97	0	1	91	18	−6	−12
1:3	−60	−97	345	−66	−69	−101	−77	−54	−57
2:1	11	0	−66	180	4	−3	108	5	−17
2:2	153	1	−69	4	219	−4	12	118	−21
2:3	1	91	−101	−3	−4	134	8	−1	−27
3:1	7	18	−77	108	12	8	174	5	−22
3:2	131	−6	−54	5	118	−1	5	215	−27
3:3	−36	−12	−57	−17	−21	−27	−22	−27	203

Таблица 3. Нормированные степени сходства s для пар белков

s	1:1	1:2	1:3	2:1	2:2	2:3	3:1	3:2	3:3
1:1	100	−4.9	−21	5.4	69	0.6	3.5	60	−17
1:2	−4.9	100	−40	0.0	0.6	67	12	−3.4	−7.0
1:3	−21	−40	100	−25	−25	−42	−30	−19	−21
2:1	5.4	0.0	−25	100	2.0	−1.9	61	2.5	−8.9
2:2	69	0.6	−25	2.0	100	−2.3	6.1	54	−10
2:3	0.6	67	−42	−1.9	−2.3	100	5.2	−0.6	−16
3:1	3.5	12	−30	61	6.1	5.2	100	2.6	−12
3:2	60	−3.4	−19	2.5	54	−0.6	2.6	100	−13
3:3	−17	−7.0	−21	−8.9	−10	−16	−12	−13	100

табл. 4 полужирным шрифтом. Сам граф показан на рис. 1.

Граф имеет три компоненты связности: две, состоящие из изолированных вершин 1:3, 3:3, и одну, содержащую все остальные вершины. Первым двум компонентам соответствуют тривиальные накрывающие деревья (из одной вершины), к ним не применима процедура разделения, и они дают два одноэлементных кластера. Рассмотрим нетривиальную компоненту связности. Для нее имеется одно накрывающее дерево D, которое получается из графа G удалением ребер, показанных на рис. 1 пунктиром. Остальные ребра в G становятся ребрами в D. Пусть параметр p равен двум. Исходное дерево D не удовлетворяет первому условию сохранения. (Если $p = 3$, то D не удовлетворяет второму условию сохранения). В D удаляется ребро 3:1–3:2. Получается набор из двух деревьев, показанный на рис. 2. Дерево с четырьмя вершинами не удовлетворяет третьему условию сохранения. Ребро 1:2–3:1 с минимальным весом удаляется. Получается набор из трех деревьев, показанный на рис. 3.

Все полученные деревья удовлетворяют сразу всем условиям сохранения. Алгоритм закончил работу. В результате по пяти деревьям получены следующие пять белковых кластеров: кластер 1 (1:1, 2:2, 3:2): {photosystem_I subunit IX [*Porphyra purpurea*], photosystem_I subunit IX [*Guillardia theta*], photosystem_I subunit IX [*Nephroselmis olivacea*]}; кластер 2 (1:2, 2:3): {photosystem_I subunit XII [*Porphyra purpurea*], photosystem_I subunit XII [*Guillardia theta*]}; кластер 3 (2:1, 3:1): {photosystem_I subunit VIII [*Guillardia theta*], photosystem_I subunit VIII [*Nephroselmis olivacea*]}; кластер 4 (1:3): {ribosomal protein S18 [*Porphyra purpurea*]}; кластер 5

Таблица 4. Граф G соответствует полужирным значениям

G	1:1	1:2	1:3	2:1	2:2	2:3	3:1	3:2	3:3
1:1				5.4	**69**	0.6	3.5	**60**	
1:2					0.6	**67**	**12**		
1:3									
2:1					**2.0**		**61**	2.5	
2:2							6.1	**54**	
2:3							5.2		
3:1								**2.6**	
3:2									
3:3									

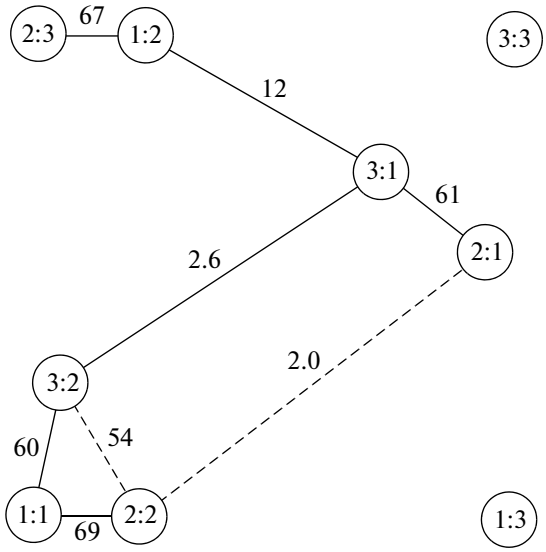

Рис. 1. Граф *G*. Длина ребра приблизительно обратно пропорциональна его весу, то есть более короткие ребра соответствуют большему сходству соответствующих белков. Изолированные вершины расположены произвольно.

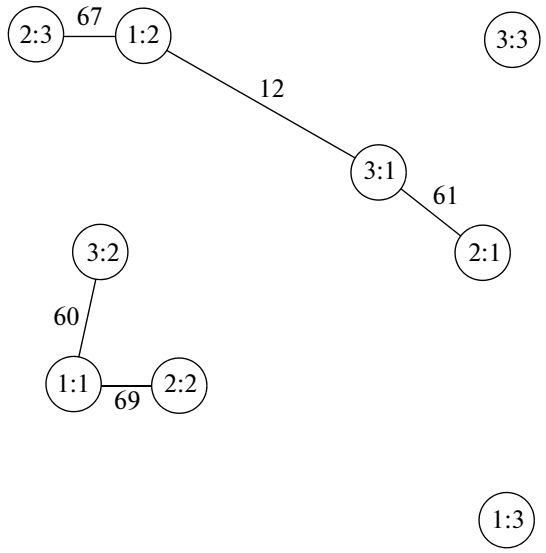

Рис. 2. Два дерева после первого разделения.

(3:3): {ribosomal protein L36 [*Nephroselmis olivacea*]}. Конец примера.

Алгоритм имеет три параметра – H, p и L. Поясним их смысл. Для полного графа белки, у которых нормированная степень сходства s не удовлетворяет неравенству $s < H$, обязательно попадают в один кластер, т.е. H – это максимально возможное сходство s между кластерами. Отметим, что два белка с нормированной степенью сходства выше H могут и не принадлежать одному ребру в дереве, но они обязательно соединены в дереве путем, каждое ребро которого соединяет белки с нормированной степенью сходства выше H. Такой путь никогда не разрезается при измельчении деревьев (будущих кластеров), поэтому эти белки обязательно попадут в один кластер. Параметр p ограничивает размер будущего кластера относительно n. Для разреженного графа параметр L ограничивает снизу степень близости двух белков; иногда имеет смысл опустить неравенство, включающее L, т.е. в разреженном графе рассматривать сколь угодно "слабые" ребра.

Приводимые ниже результаты получены в случае разреженного графа, в качестве значений параметров использовали $H = 0.7$, $p = 2$, $L = 0$, тот же результат сохраняется в диапазоне значений $0.6 \le H \le 0.7$, $1 \le p$ и $L \le 0.05$. Несколько слов о влиянии значений параметров: при $p < 1$ кластеры максимального размера разрушаются; при неограниченном времени счета можно брать большие значения p, даже $p = +\infty$, т.е. опускать условие сохранения 1. Если значение L превышает 0.05, то число ребер начинает быстро уменьшаться, а число компонент связности быстро возрастать, при этом разрушаются "слабые" кластеры. При небольшом увеличении значения L изменения в кластерах носят скорее положительный характер. При $H \le 0.55$ некоторые кластеры объединяются, а при $H \ge 0.75$ – разрушаются.

Отметим, что алгоритм кластеризации может выдавать несколько вариантов: при построении накрывающего графа D несколько таких графов; при каждом увеличении глубины дерева $D_{i,j}$: несколько вариантов новой пары деревьев. Для рассмотренных нами видов число всех этих вариантов мало (см. следующий абзац). Это позволило нам применять алгоритм, выбирая каждый раз только один из вариантов. При этом могут возникать трудности, которые преодолеваются вручную, за счет дополнительной биологической информации. Например, кластер L-субъединиц протохлорофиллидредуктазы ChlL был вручную выделен из большего кластера, сформированного алгоритмом и объединяющего различные белки, которые не встречаются совместно ни в одном пластоме. Эволюция генов *chlB*, *chlL* и *chlN*, кодирующих субъединицы независимой от света протохлорофиллидредуктазы, описана ранее [22]. Вручную выделены еще два кластера, один из которых составляет фрагменты β"-субъединицы РНК-полимеразы бактериального типа у Piroplasmida (*Babesia bovis* и *Theileria parva*), а другой – киназы из водорослей *Rhodomonas salina* и *Heterosigma akashiwo*.

При построении дерева D конкурентами можно считать ребра веса s, которые не входят в D при том, что некоторое другое ребро с этим весом в него входит. Таких случаев насчитывается несколько сотен – из общего числа ребер порядка

ста тысяч в разреженном графе. На обсуждаемом наборе данных процедура разделения деревьев однозначна. Вообще говоря, учет альтернативных вариантов сократит ручную работу и может быть полезен.

На основе алгоритма кластеризации создана база данных для поиска белков по филогенетическому профилю и других исследований пластомов (доступна по адресу [19]). Результаты, полученные с помощью этой базы, и сам алгоритм доложены на 53-й и 54-й Научных конференциях МФТИ, на юбилейной биологической конференции, посвященной 50-летию ИППИ РАН [8, 23, 24].

Для контроля наших результатов и построения филогенетических деревьев при исследовании РНК-полимераз использовали пакет программ MEGA 5 [25]. Поиск субъединиц РНК-полимераз выполнялся программой BLAST [26], соответствующее ожидаемое значение (e-value) ниже обозначается E.

Сайт [19] обеспечивает, по крайней мере, две функции: поиск белка по филогенетическому профилю и поиск по фрагменту аминокислотной последовательности всех белков, кодируемых в пластидах родофитной ветви и содержащих данный фрагмент, и кластеров этих белков. Там же приведены инструкция пользователя и примеры вычислений.

МАТЕРИАЛЫ ИССЛЕДОВАНИЯ

Пластомы, указанные в табл. 1, получены из базы данных NCBI. В их числе, пластомы недавно секвенированных диатомовых водорослей [27, 28]. Некоторые фрагменты ядерных геномов *Eimeria tenella* и *Neospora caninum* Liverpool получены из базы данных Sanger Institute [29].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Кластеризация пластомных белков

Мы рассмотрели многочисленные таксономические группы родофитной ветви, которые охватывают *все* виды этой ветви, представленные в базе данных GenBank NCBI (табл. 1). Рассмотрено 3426 белков, из них образовано 260 кластеров, содержащих строго более одного белка, и 143 одноэлементных кластера. Последние в совокупности содержат 4% от числа всех белков, 11 кластеров состоят целиком из паралогичных друг другу белков. В большинстве кластеров паралоги отсутствуют: 359 кластеров не содержат паралогов и 44 кластера содержат их. Распределение кластеров по числу представленных в них видов показано на рис. 4.

Удалось выделить белки, которые характеризуют несколько таксономических групп, т.е. они кодируются в их пластомах и только в них. Эти белки

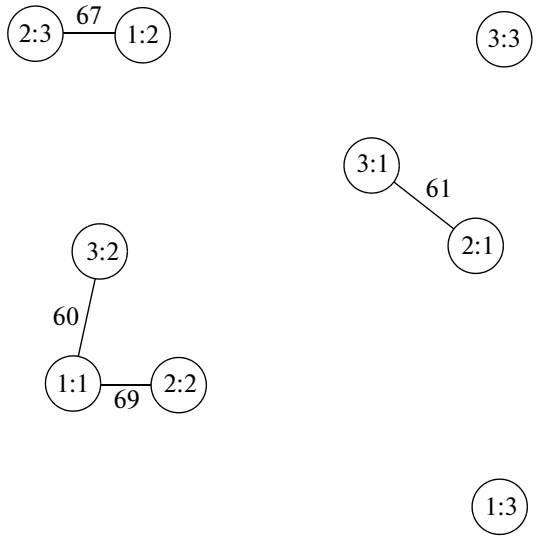

Рис. 3. Набор деревьев после второго шага разделения.

и группы перечисляются ниже. Белки, общие для пластомов всех рассмотренных видов, составляют 8 кластеров: рибосомные белки S2, S12, L2, L6, L14 и L16, фактор элонгации Tu и β-субъединица РНК-полимеразы бактериального типа. Рибосомный белок S19 определен у всех рассмотренных видов, кроме споровика *Babesia bovis*.

Белки, кодируемые в пластидах багрянок (*Cyanidioschyzon merolae*, *Cyanidium caldarium*, *Gracilaria tenuistipitata*, *Porphyra purpurea* и *P. yezoensis*), но не в остальных рассмотренных пластомах, образуют 24 кластера: третий фактор инициации трансляции; α-, β-, β_{18}-, γ-субъединицы аллофикоцианина; α- и β-субъединицы фикоцианина; два формообразующих белка фикобилисом и связанный с деградацией фикоболисом белок Ycf18; тиоредоксин; белки комплекса ацетил-CoA-карбоксилазы; пренилтрансфераза; ацетилглутаматкиназа; ферредоксин-зависимая глутаматсинтаза; α- и β-субъединицы пируватдегидрогеназы E1; субъединицы антранилатсинтазы; α-субъединица триптофансинтазы; и гипотетические консервативные белки.

Не найдено белка, который был бы специфичен для криптофитовых водорослей *Cryptomonas paramecium*, *Guillardia theta* и *Rhodomonas salina,* а также – специфичного для Chromerida (*Alveolata* sp. CCMP3155 и *Chromera velia*).

Белки, специфичные для споровиков группы Piroplasmida (*Babesia bovis*, *Theileria parva*), составляют пять кластеров: два из них – слабые гомологи рибосомных белков, еще два – молекулярные шапероны, гомологичные ClpC (это: YP_002290851.1, XP_762692.1, YP_002290850.1, XP_762693.1) и фрагменты β"-субъединицы РНК-

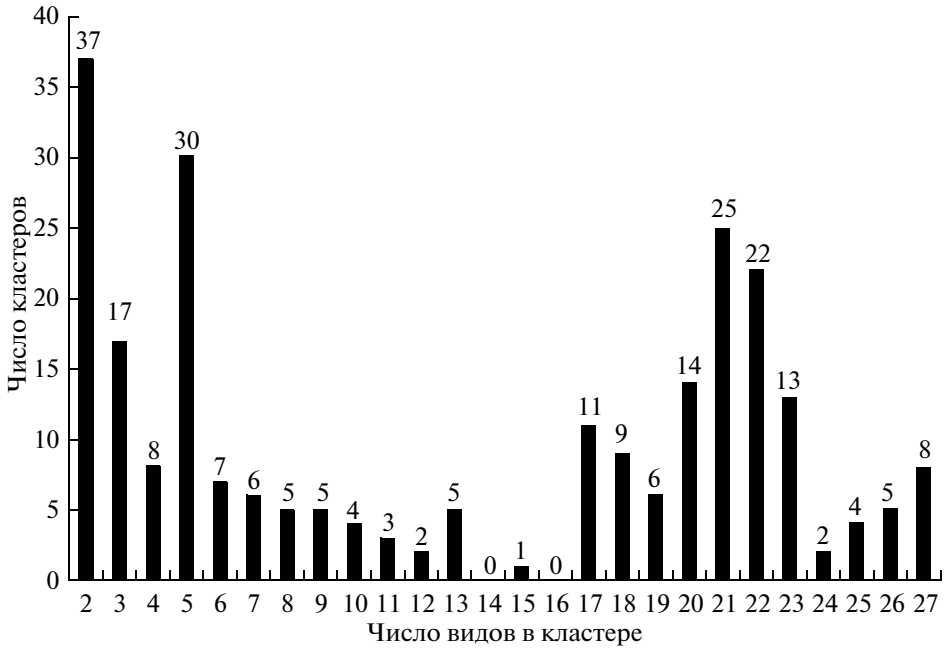

Рис. 4. Распределение кластеров по числу видов. Показано число кластеров пластомных белков в зависимости от числа представленных в кластере видов родофитной ветви; 154 кластера содержат белки только из одного вида.

полимеразы бактериального типа (YP_002290845.1, XP_762712.1).

Группа "Diatoms и Dinotoms" содержит *Durinskia baltica*, *Kryptoperidinium foliaceum*, *Fistulifera* sp. JPCC DA0580, *Odontella sinensis*, *Phaeodactylum tricornutum*, *Thalassiosira oceanica*, *Thalassiosira pseudonana*. Среди них пять пластомов диатомовых водорослей: *Fistulifera* sp. JPCC DA0580, *P. tricornutum*, *O. sinensis, T. oceanica* и *T. pseudonana*. Специфичными для этой группы являются два кластера белков: один содержит гомологи белка Ycf88, другой представлен парами паралогов, гомологичных белку Ycf89.

Поиск РНК-полимераз в ядерных геномах споровиков

У штаммов *Toxoplasma gondii* ME49 (XP_002367014.1), *T. gondii* VEG (EEE31947.1), *T. gondii* GT1 (EEE23737.1) и у *Neospora caninum* (CBZ55882.1) найдено по одному экземпляру РНК-полимеразы фагового типа с номерами, указанными в скобках. При этом белки штаммов *T. gondii* ME49 и VEG совпадают, а белок штамма GT1 содержит замены аминокислотных остатков в нескольких позициях и вставку в позициях от 347 до 354. У *Eimeria tenella* не удалось определить РНК-полимеразу фагового типа.

Гомологи РНК-полимераз фагового типа найдены у многих споровиков, не являющихся кокцидиями: *Plasmodium berghei* (XP_676913.1), *Pl. falciparum* 3D7 (XP_001347935.1), *Pl. knowlesi* H (XP_002259256.1), *Pl. vivax* SaI-1 (XP_001615369.1), *Pl. yoelii* 17XNL (XP_727223.1), *Pl. chabaudi* (XP_739650.1), *Babesia bovis* (XP_001611431.1), *Theileria annulata* (XP_953797.1), *Th. parva* (XP_766496.1). Дерево РНК-полимераз фагового типа показано на рис. 5. Однако не найден ортологичный белок у кокцидии *Cryptosporidium parvum*, которая, в отличие от многих споровиков, не имеет пластид.

В ядерном геноме *Toxoplasma gondii* обнаружен только один ген, кодирующий σ-субъединицу РНК-полимеразы. Ее длина – 1002 а.о. у штаммов ME49 и GT1, 1001 а.о. – у штамма VEG. Ниже рассматривается белок XP_002367841.1 штамма ME49. В ядерном геноме *Neospora caninum* ген CBZ51366.1 кодирует σ-субъединицу РНК-полимеразы длиной 1206 а.о. С-концы σ-субъединиц РНК-полимераз у *T. gondii* и *N. caninum* чрезвычайно сходны друг с другом, однако не имеют существенного сходства с σ-субъединицами диатомовых водорослей *Phaeodactylum tricornutum* CCAP 1055/1 и *Thalassiosira pseudonana* CCMP1335, золотистой водоросли *Aureococcus anophagefferens*, криптофитовых водорослей *Guillardia theta* и *Hemiselmis andersenii*. У кокцидий σ-субъединицы, ближайшие к этим σ-субъединицам, найдены у цианобактерий *Cyanothece* sp. PCC 7822 (YP_003885480.1), *Microcoleus chthonoplastes* PCC 7420 (ZP_05024793.1), *Acaryochloris marina* MBIC11017 (YP_001519047.1) и у δ-протеобактерии *Desulfarculus baarsii* DSM 2075 (YP_003809216.1). Бактериальные ортологи имеют длины от 260 до 363 а.о. У всех видов хоро-

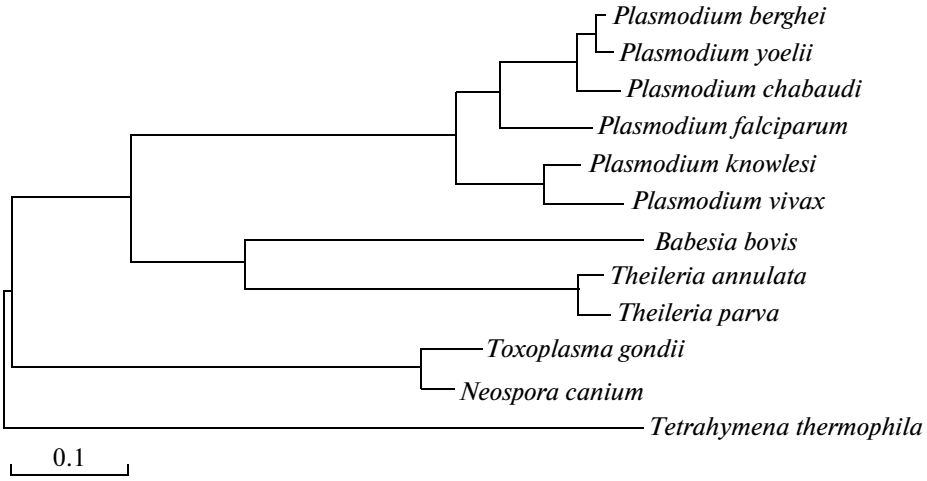

Рис. 5. Дерево РНК-полимераз фагового типа у простейших надтипа Alveolata.

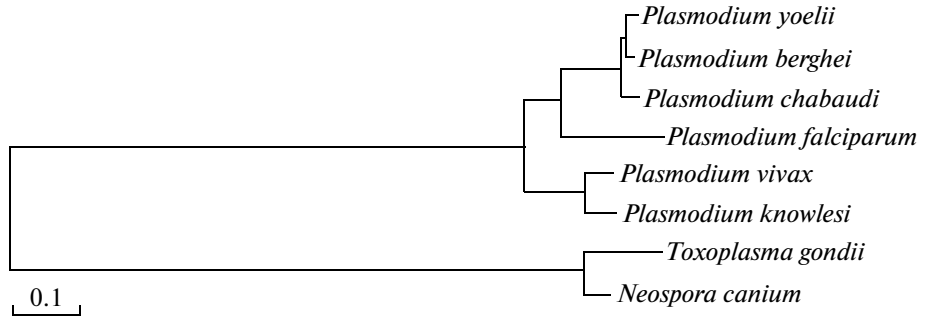

Рис. 6. Дерево σ-субъединиц РНК-полимераз у споровиков.

шо выравниваются С-концы второго региона, весь третий регион и N-концы четвертого региона σ-субъединиц РНК-полимераз. Четвертый регион *T. gondii, N. caninum* и *D. Baarsii* выравнивается по всей длине.

Ортологи σ-субъединиц РНК-полимеразы найдены также у простейших из отряда Haemosporida: *Plasmodium berghei* (XM_669238.1), *Pl. falciparum* 3D7 (XP_966194.1), *Pl. knowlesi* H (XM_002261430.1), *Pl. vivax* SaI-1 (XP_001616222.1), *Pl. yoelii* 17XNL (XP_724777.1), *Pl. chabaudi* (XM_739944.1). В каждом из них отсутствуют другие σ-субъединицы. Не удалось определить σ-субъединицы РНК-полимеразы у видов из отряда Piroplasmida: *Theileria parva, Th. annulata, Babesia bovis*. Дерево σ-субъединиц показано на рис. 6.

Особенностью пластомов споровиков является отсутствие у них α-субъединиц РНК-полимераз бактериального типа. Мы рассмотрели три вида кокцидий: *Eimeria tenella*, *Toxoplasma gondii* и *Neospora caninum*. Данные о *T. gondii* и об обсуждаемых водорослях и бактериях доступны в базе данных NCBI. У *T. gondii* ME49 α-субъединица кодируется в ядре, соответствующий белок XP_002367289.1 имеет 836 а.о. Этот белок отличается по одной позиции в штаммах *T. gondii* ME49 и GT1. В ядерном геноме *E. tenella* обнаружена близкая ($E = 1.1 \times 10^{-71}$) α-субъединица, для которой определены фрагменты четырех экзонов на контиге dev_EIMER_contig_00028796 с координатами соответственно 5283..5453, 5682..6167, 6576..6785 и 7273..7965. В ядерном геноме *N. caninum* обнаружена близкая ($E = 9.9 \times 10^{-288}$) α-субъединица, в которой имеются два экзона на контиге Contig892 с координатами соответственно 45655..47412 и 47940..48611.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Как указано в разделе “Результаты исследования”, большинство наиболее консервативных белков связано с трансляцией.

Белок NP_045121.1 у *Cyanidium caldarium* входит в кластер, содержащий белки YP_537023.1 из *Porphyra yezoensis* и NP_053952.1 из *P. purpurea*. Эти белки имеют относительно короткий консервативный домен, характерный для транскрипционного фактора NtcA (Ycf28). Белок NP_849012.1

из *Cyanidioschyzon merolae* является гомологом NtcA, однако он не вошел в кластер NtcA из-за значительного отличия, в том числе, в наиболее консервативном домене фактора. Еще меньше сходства в соответствующем домене NtcA и его гомологе у *Gracilaria tenuistipitata*. Эволюционно это изменение связано с элиминацией из пластома гена *glnB*, транскрипция которого регулируется фактором NtcA у багрянок *Porphyra* spp. и *Cyanidium caldarium* [30].

Пластом *Gracilaria tenuistipitata* содержит гены *leuC* и *leuD*, кодирующие большую (YP_063540.1) и малую (YP_063541.1) субъединицы 3-изопропилмалатдегидрогеназы, которые отсутствуют в других рассмотренных пластомах. Как отмечалось ранее [31], это свидетельствует о раннем разделении таксономических групп Florideophyceae (включающей *G. tenuistipitata*) и Bangiophyceae в составе отдела багрянок.

Особенностью пластомов споровиков является отсутствие в них α-субъединиц РНК-полимераз бактериального типа, однако в этих пластомах найдены их гомологи в ядерных геномах большинства споровиков.

Наличие у диатомовых водорослей и близких к ним третичных эндосимбионтов общих белков, отсутствующих в пластидах других видов, позволяет предположить, что диатомовые водоросли обособились от других представителей родофитной ветви, ранее других.

Неконсервативность большинства субъединиц РНК-полимеразы бактериального типа у Piroplasmida позволяет сомневаться в работоспособности этого фермента, поскольку в их ядерных геномах не удалось определить σ-субъединицу. По-видимому, у Piroplasmida транскрипция всего пластома осуществляется исключительно РНК-полимеразами фагового типа. Это означает, что применение в борьбе с Piroplasmida антибиотиков, ингибирующих РНК-полимеразу бактериального типа, неэффективно. Напротив, такие антибиотики могут использоваться против *Plasmodium* spp, *Toxoplasma gondii* и *Neospora caninum*.

Дерево σ-субъединиц РНК-полимераз бактериального типа у споровиков, исключая виды из Piroplasmida, хорошо согласуется и с деревом видов, и с деревом РНК-полимераз фагового типа. Тот факт, что у споровиков существует не более одной σ-субъединицы РНК-полимеразы, указывает на незначительную роль регуляции пластома на уровне транскрипции. Вероятно, наиболее важна в этом случае регуляция на уровне трансляции или процессинга, что подтверждается другими наблюдениями [7].

РНК-полимеразы фагового типа у видов рода *Plasmodium* хорошо выравниваются между собой, образуя кладу на дереве белков; эти полимеразы формируют также отдельные клады Piroplasmida и Coccidia. Однако РНК-полимеразы Coccidia существенно отличаются от ортологичных белков других споровиков. Напротив, РНК-полимеразы фагового типа у кокцидий близки к ортологичным белкам тетрахимены, не имеющей пластид. Можно предположить, что у кокцидий РНК-полимеразы фагового типа не играют роли в транскрипции пластома. Мы не обнаружили скольнибудь существенного разнообразия РНК-полимераз фагового типа у простейших. Вероятно, РНК-полимеразы фагового типа у споровиков имеют древнее происхождение и не связаны с приобретением пластид. Напротив, у высших растений наблюдается большое разнообразие РНК-полимераз фагового типа, которые нацелены на различные органеллы [11, 8].

Сходные результаты поучены нами при изучении хлорофитной ветви растений и водорослей, которые будут опубликованы.

Работа получила финансовую поддержку Государственных контрактов (14.740.11.0624, 14.740.11.1053, 14.740.12.0830) Министерства образования и науки РФ.

СПИСОК ЛИТЕРАТУРЫ

1. Lemieux C., Otis C., Turmel M. 2007. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biology*. **5**, 1–17.
2. Imanian B., Pombert J.-F., Keeling P.J. 2010. The complete plastid genomes of the two 'Dinotoms' *Durinskia baltica* and *Kryptoperidinium foliaceum*. *PLoS ONE*, **5**(5), e10711.
3. Балашов Ю.С. 1998. *Иксодовые клещи – паразиты и переносчики инфекций*. СПб.: Наука.
4. Brayton K.A., Lau A.O.T., Herndon D.R., Hannick L., Kappmeyer L.S., et al. 2007. Genome sequence of *Babesia bovis* and comparative analysis of Apicomplexan Hemoprotozoa. *PLoS Pathogens*. **3**, e148.
5. Wilson R.J.M., Rangachari K., Saldanha J.W., Rickman L., Buxton R.S., Eccleston J.F. 2003. Parasite plastids: maintenance and functions. *Phil. Trans. R. Soc. Lond*. **B**. **358**, 155–164.
6. Zhu G., Marchewka M.J., Keithly J.S. 2000. *Cryptosporidium parvum* appears to lack a plastid genome. *Microbiol.* **146**, 315–321.
7. Садовская Т.А., Селиверстов А.В. 2009. Анализ 5'-лидерных областей некоторых генов пластид у простейших типа Apicomplexa и у красных водорослей. *Молекулярная биология*, **43**, 599–604.
8. Селиверстов А.В., Любецкий В.А. 2011. Эволюция РНК-полимераз и их промоторов в пластидах. *Юбилейная конференция 50 лет ИППИ РАН*. Москва. С. 58–62.
9. Jeruzalmi D., Steitz T.A. 1998. Structure of T7 RNA polymerase complexed to the transcriptional inhibitor T7 lysozyme. *EMBO J*. **17**, 4101–4113.

10. Ma N., McAllister W.T. 2009. In a head-on collision, two RNA polymerases approaching one another on the same DNA may pass by one another. *J. Mol. Biol*. **391**, 808–812.
11. Kühn K., Bohne A.-V., Liere K., Weihe A., Thomas Börner T. 2007. *Arabidopsis* phage-type RNA polymerases: accurate in vitro transcription of organellar genes. *Plant Cell*. **19**, 959–971.
12. Altenhoff A.M., Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* **5**: e1000262.
13. Интернет-ресурс http://orthomcl.cbil.upenn.edu/
14. Интернет-ресурс http://roundup.hms.harvard.edu/browse/
15. Интернет-ресурс http://www.omabrowser.org/
16. Интернет-ресурс http://eggnog.embl.de/
17. Интернет-ресурс http://inparanoid.sbc.su.se/
18. Интернет-ресурс http://www.ncbi.nlm.nih.gov/COG/
19. Интернет-ресурс http://lab6.iitp.ru/ppc/redline/
20. Needleman S.B., Wunsch C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*. **48**, 443–453.
21. Интернет-ресурс http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt
22. Fong A., Archibald J.M. 2008. Evolutionary dynamics of light-independent protochlorophyllide oxidoreductase genes in the secondary plastids of Cryptophyte algae. *Eukaryotic Cell*. **7**, 550–553.
23. Зверков О.А., Селиверстов А.В., Любецкий В.А. 2010. Об одном алгоритме кластеризации белков. *Труды 53-й научной конференции МФТИ*. М.: МФТИ, часть 1, Т. 1, С. 118–119.
24. Зверков О.А., Горбунов К.Ю., Селиверстов А.В., Любецкий В.А. 2011. Кластеризация белков с учетом их доменной структуры. *Труды 54-й научной конференции МФТИ "Проблемы фундаментальных и прикладных естественных и технических наук в современном информационном обществе". Управление и прикладная математика*. Том 2. М.: МФТИ, С. 88–89.
25. Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol*. **28**, 2731–2739.
26. Интернет-ресурс http://blast.ncbi.nlm.nih.gov/
27. Lommer M., Roy A.-S., Schilhabel M., Schreiber S., Rosenstiel P., LaRoche J. 2010. Recent transfer of an iron-regulated gene from the plastid to the nuclear genome in an oceanic diatom adapted to chronic iron limitation. *BMC Genomics*. **11**, 13.
28. Tanaka T., Fukuda Y., Yoshino T., Maeda Y., Muto M., Matsumoto M., Mayama S., Matsunaga T. 2011. High-throughput pyrosequencing of the chloroplast genome of a highly neutral-lipid-producing marine pennate diatom, *Fistulifera* sp. strain JPCC DA0580. *Photosynthesis Res.***109**, 223–229.
29. Интернет-ресурс http://www.sanger.ac.uk/
30. Лопатовская К.В., Селиверстов А.В., Любецкий В.А. 2011. Регулоны NtcA и NtcB у цианобактерий и хлоропластов водорослей отдела Rhodophyta. *Молекуляр. биология*, **45**, 570–574.
31. Hagopian J.C., Reis M., Kitajima J.P., Bhattacharya D., de Oliveira M.C. 2004. Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. liui provides insights into the evolution of rhodoplasts and their relationship to other plastids. *J. Mol. Evol*. **59**, 464–477.