ISSN 0026-8933, Molecular Biology, 2012, Vol. 46, No. 5, pp. 717–726. © Pleiades Publishing, Inc., 2012.
Original Russian Text © O.A. Zverkov, A.V. Seliverstov, V.A. Lyubetsky, 2012, published in Molekulyarnaya Biologiya, 2012, Vol. 46, No. 5, pp. 799–809.

BIOINFORMATICS

UDC 577.21+519.163

# Plastid-Encoded Protein Families Specific for Narrow Taxonomic Groups of Algae and Protozoa

**O. A. Zverkov, A. V. Seliverstov, and V. A. Lyubetsky**

*Kharkevitch Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia;*
*e-mail: zverkov@iitp.ru*

Received December 1, 2011; in final form March 2, 2012

**Abstract**—Protein clustering is useful for refining protein annotations and searching for proteins by their phylogenetic profile. We have performed the clustering of proteins encoded in the plastoms of Rhodophyta, as well as other plastid-containing species related to the Rhodophyta branch. The corresponding database and cluster search according to protein phylogenetic profile are available at http://lab6.iitp.ru/ppc/redline. Plastome-encoded proteins specific for small taxonomic groups of algae and protozoa have been found based on this database, and the search for and analysis of RNA polymerase in the nuclear genomes of Apicomplexa has been performed.

**DOI:** 10.1134/S0026893312050123

*Keywords*: algae, Apicomplexa, plastids, Rhodophyta branch, protein families, protein clusters, phylogenetic profile

## INTRODUCTION

The classification of proteins of any taxonomic group into families, i.e., protein clustering, is useful to refine protein annotation and search for the protein based on its phylogenetic profile. For example, protein clustering allows one to judge the functional ability of multisubunit protein complexes, e.g., bacterial-type RNA polymerases. For this purpose, the presence of all necessary subunits must be verified for a given species based on the cluster composition for each subunit. In addition, clustering provides insights into plastom evolution.

One should note that the phylogenetic profile of a protein is represented by +1 when the protein is represented in species $x$, and −1 when the protein is not represented in species $x$, for all species $x$ from some set $X$. These signs (numbers) are oriented along a vector which positions are marked by the index $x$ and the set $X$ is ordered by a random fixed mode.

We consider plastids of Rhodophyta and other Rhodophyta plastid related species. All these species form the Rhodophyta branch in the tree of plastids [1] and will be called so. The list of plastoms considered is given in Table 1. In particular, diatomaceous algae are of major interest; their five plastoms and two complete nuclear genomes are known. Along with diatomaceous algae, we considered two representatives of the Alveolata supertype, i.e., *Durinskia baltica* (NC_014287.1) and *Kryptoperidinium foliaceum* (NC_014267.1), the plastoms of which are fully sequenced and closely related to the plastom of *Phaeodactylum tricornutum* [2].

The Rhodophyta branch of plastids includes apicoplasts of many Apicomplexa, i.e., organelles that resemble plastids of Rhodophyta but have a highly reduced genome. The study of Apicomplexa is of special interest, since they cause human and animal diseases. In particular, *Theileria* and *Babesia* are transmitted by ixodes [3] and cause the following diseases in cattle: *B. bigemina* and *B. bovis* cause babesiosis of cattle; *Th. annulata* causes theileriosis of cattle; *Th. parva* causes East Coast fever; *Eimeria tenella* causes eimeriosis in hens; and *Toxoplasma gondii* causes toxoplasmosis of cats and humans. Various species of *Plasmodium* cause malaria in humans (*Pl. falciparum*), rodents, and other animals. The genomes of *B. bovis* and *Th. parva* are extremely closely related to each other [4]. The peculiarities and functions of apicoplasts are reviewed in [5]. We note that some Apicomplexa, e.g., *Cryptosporidium parvum* [6] do not carry plastids.

The investigation of various processes associated with apicoplasts will facilitate the understanding of their role in the transmission of infection and the mechanisms of the drug's action on apicoplast. Since translation and, usually, transcriptions are bacterial in nature in apicoplasts, they serve the main target of antibiotics that have no direct effect on the expression of nuclear and mitochondrial genes; therefore, the investigation of regulatory mechanisms and the evolution of these processes is urgent. Some results in this area are given in our previous studies [7, 8].

Since many proteins that reach plastids are encoded in the nucleus, it is necessary to match data on proteins formed in the nucleus with data on the

**Table 1.** Plastoms of Rhodophyta branch

| Locus number | Species | Number of proteins | Quantity of clusters | |
|---|---|---|---|---|
| | | | >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 |

Note: First column indicates the number of plastom in the NCBI database, second column indicates species to which the plastom belongs, third column is the number of plastom proteins in this species, fourth column is the number of plastom protein clusters for all species indicated in the table, including the members of this species with a number of proteins strictly greater than and equal to 1, respectively.

genes and regulatory regions in plastome. The subunits of bacterial-type RNA polymerase and phage-type RNA polymerases homologous to the RNA polymerases of the bacteriophage T7 [9, 10], which are responsible for transcription in plastids and mitochondria, are of special significance [11].

Protein clustering is an independent task that is useful for refining protein annotation, and conducting an effective search for protein according to its phylogenetic profile and estimating the organism's potential to adapt to various conditions. In particular, a database based on clustering will be useful for solving the aforementioned tasks. Several such databases [12] are known; nevertheless, the majority of them contain a small number of species, of which only a few carry plastids of the Rhodophyta branch. For example, the OrthoMCL database [13] (last version of March 31, 2011) includes 150 genomes, and only a few of them belong to the Rhodophyta branch which we considered; the RoundUp [14] database contains a small number of species, which includes only a few algae and Apicomplexa; the OMA database [15] (version of May 18, 2011) covers 1109 species, though hardly any of them contain plastids from the Rhodophyta branch; the current version of the EggNOG database [16] includes 1133 species, but it does not present species of the Rhodophyta branch; the InParanoid [17] (version 7.0) contains only 100 eukaryotic organisms, among which only one diatomaceous alga and several Apicomplexa belong to the Rhodophyta branch; and the COG and

KOG [18] databases present a small number of species, including only two plants and no species of the considered Rhodophyta branch.

Plastid encoded protein clustering will create a novel database that is particularly suitable for studying Apicomplexa, i.e., the causative agents of many protozoan infections.

We have performed the plastome-encoded protein clustering of the Rhodophyta branch. Part of this database is presented in the supplement (for additional material, go to www.molecbio.com/downloads/2012/5/supp_zverkov_eng/pdf). A cluster search in this database based on the protein phylogenetic profile is available on the site [19]. Plastome-encoded proteins specific for small taxonomic groups of algae and protozoa, as well as RNA polymerases in the nuclear genomes of Apicomplexa and, in particular, σ-subunits of bacterial-type RNA polymerase and phage-type RNA polymerase in species of the Alveolata supertype have been found based on this database.

## ALGORITHM OF PROTEIN FAMILY CLUSTERING

We propose the following clustering algorithm of protein amino acid sequences (words of various length in 20 characters) based on their similarity. In particular, the algorithm was applied to investigate algae and Apicomplexa and generate the plastid protein family database.

The clusters are formed by decomposition starting from a single cluster that contains all of the proteins. The cluster may include distant proteins if, during decomposition, they did not fall into various clusters. This approach is useful when considering distant species and their proteins that originated from one ancestor protein and maintained common functions when the similarity of these proteins is comparable and there is less similarity between many paralogs (close homologs from one genome). In addition, our algorithm operates very quickly in quadratic time with regard to the total number of species $n$. Let us consider the following algorithm.

Let a set of plastids be given that are designated by index $i$ and, for each plastid, their proteins are given $P_{ij}$ and designated with the index $j$. For all pairs of proteins $(P_{ij}, P_{kl})$ for all pairs of species $(S_i, S_k)$, the closeness characteristic $s_0(P_{ij}, P_{kl})$ is calculated for proteins as the optimal global alignment quality of these sequences; the pair alignment is not used and does not have to be calculated. This characteristic is calculated using the standard Needleman–Wunsch algorithm [20], in which the sum of corresponding elements of the BLOSUM62 matrix is used as the sequence similarity measure [21]. Then, the algorithm calculates the normalized degree of similarity $s(P_{ij}, P_{kl})$ of proteins using the formula $2s_0(P_{ij}, P_{kl})(s_0(P_{ij}, P_{ij}) + s_0(P_{kl}, P_{kl}))^{-1}$.

The complete undirected graph $G_0$ is considered with a vertex set $\{P_{ij}\}$, in which each edge $(P_{ij}, P_{kl})$ is assigned a value $s(P_{ij}, P_{kl})$,, which we will call the edge weight. The edges connect various vertices, i.e., loops are absent.

In order to narrow the scope of calculations, sparse graph $G$ may be used instead of the complete graph, which consists only of edges $(P_{ij}, P_{kl})$, that satisfy the followingconditions:

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

in which the maximums are taken for all proteins from the corresponding $i$ and $k$ characters and $L$ is algorithm parameter, which is 0 by default. If $i = k$, then one more condition is supposed, i.e., $m \neq l$, and the second equality may be omitted.

In the obtained graph $G$, the algorithm creates a set of all of its connected components. $G$ covering unrooted trees $D$ are then constructed for each component (designated with the same character $G$). The latter refers to the following: in $G$, edges are walked in the order of their decreasing weight, which are called the edges of the constructed tree $D$; if the addition of the current edge from $G$ to $D$ leads to the appearance of a cycle in $D$, this edge is left out. As a result, $D$ does not contain cycles, i.e., it is a tree and includes all the vertices from $G$. The sum of edge weights in $D$ is called the weight of the tree $D$; the obtained trees have the maximum possible weight. Thus, for each connected component in the initial graph $G$, trees $D$ covering the component are constructed.

The following recursive "division" procedure is then applied to each tree $D$, which constructs the set of trees $\{D_{i,j,..}\}$, in which all the indices are represented by 1 or 2. The length of the $i, j, ..,$ index sequence is called the depth of a tree $D_{i,j,..}$. If the current tree, e.g., $D_{i,j,..}$ of some depth $k$ of this set, does not satisfy the criterion of tree division formulated below, then it is replaced in the set by two trees $D_{i,j,...,1}$ and $D_{i,j,...,2}$ of depth $k + 1$ each, by means of removing $e_0$ edge from $D_{i,j,..}$ with minimal weight $s$ along the entire $D_{i,j,..}$ when the condition $s < H$ has been satisfied, where $H$ is the algorithm parameter (if this inequality is not satisfied, the current tree is not divided and we go to the next tree). Otherwise, the criterion of maintaining the current tree $D_{i,j,...}$ without a change is verified. This "conservation" criterion for the $D_{i,j,..}$ tree with vertex set $V$ consists of the fulfillment of three conditions as follows:

(1) $|V| \leq pn$, where $|V|$ is the number of vertices in the $D_{i,j,..}$ tree, $n$ is the number of all species in the initial set of species, and $p$ is algorithm parameter;

(2) the edge $(P_{mq}, P_{kl})$ with minimal weight in the $D_{i,j,..}$ tree connects the $P_{mq}$ and $P_{kl}$ proteins, where the indices are $m \neq k$;

**Table 2.** Degree of similarity $s_0$ for pairs of proteins

| $s_0$ | 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 |

(3) any pair of vertices $P_{mq}$ and $P_{ml}$ of the $D_{i,j,..}$ tree, which correspond to the proteins of one species, is connected in $D_{i,j,..}$ by a path that consists of vertices that correspond to proteins of the same species. If this criterion is fulfilled and there is a next tree, we go to this tree. If all of the trees have been completed, the algorithm completes its operation.

As a result, the obtained set of trees represents decomposition of the initial proteins into clusters which consist from sequences assigned to all the vertices of one tree.

### Artificial Example to Illustrate the Algorithm Operation

**Initial data.** Randomly taken proteins encoded in three plastoms are clustered: NC_000925 (*Porphyra purpurea*), NC_000926 (*Guillardia theta*), NC_000927 (*Nephroselmis olivacea*). Namely, three short proteins were taken from each plastome:

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.

Let us designate these proteins in the indicated order as 1:1, 1:2, 1:3; 2:1, 2:2, 2:3 : 3:1, 3:2, 3:3. Thus, the *n*:*m* pair indicates *m* protein of *n* plastom. The degree of similarity $s_0$ is given for all pairs of proteins in Table 2; the lower triangular part of this table coincides with the upper, and its diagonal is not used. The normalized degrees of similarity *s* for all pairs of proteins are given in Table 3. Numerical values are rounded to the nearest hundredth.

Set *L* threshold equal to null. After removing the edges in graph *G*, 15 numbers are left in Table 4 according to the second condition in the formula (1). After removing edges in graph *G* according to the first condition in the formula (1), 8 numbers are left indicated in Table 4 by bold style. The graph itself is shown in Fig. 1. The graph has three connected components, two of which consist of isolated vertices 1:3 and 3:3, and one that contains all of the other vertices. Trivial covering trees correspond to the first two components (of one vertex), the devision procedure is not applicable to them, and they produce two one-element clusters. Let us consider the nontrivial connected component. One covering tree *D* is available for this component, which is produced from graph *G* through edge removal, as shown in Fig. 1 in a dotted line. The remaining edges in *G* became edges in *D*. Let the parameter *p* equal two. The initial tree *D* does not satisfy the first conservation condition. (If $p = 3$, then *D* does not satisfy the second condition). In *D*, the edge

**Table 3.** Normalized degree of similarity *s* for pairs of proteins

| *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 |

3:1–3:2 is removed. Figure 2 shows the obtained set of two trees. The tree with four vertices does not satisfy the third conservation condition. The edge 1:2–3:1 with minimal weight is removed. A set of three trees is obtained shown in Fig. 3.

All constructed trees simultaneously satisfy all conservation conditions. The algorithm finishes its operation. As a result, the following five protein clusters were obtained for five trees:

cluster 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*]}; cluster 2 (1:2, 2:3): {photosystem_I subunit XII [*Porphyra purpurea*], photosystem_I subunit XII [*Guillardia theta*]}; cluster 3 (2:1, 3:1): photosystem_I subunit VIII [*Guillardia theta*], photosystem_I subunit VIII [*Nephroselmis olvacea*]}; cluster 4 (1:3): {ribosomal protein S18 [*Porphyra purpurea*]}; cluster 5 (3:3): {ribosomal protein L36 [*Nephroselmis olivacea*]}. End of the example.

The algorithm has three parameters: $H$, $p$, and $L$. Let us explain their meaning. For a complete graph, proteins whose normalized degree of similarity $s$ do not satisfy the inequality $s < H$, surely fall into one cluster, i.e., $H$ is the maximally possible similarity $s$ between clusters. It should be noted that two proteins with normalized degrees of similarity higher than $H$ may also not belong to one edge in the tree, but they are necessarily connected in a tree by a path, each edge of which connects proteins with normalized degree of similarity higher than $H$. This path is never cut during the division of the tree (into future clusters); therefore, these proteins surely to fall into one cluster. Parameter $p$ restricts the size of the future cluster relative to $n$. For a sparse graph, parameter $L$ restricts the degree of similarity of two proteins; it is sometimes reasonable to leave out the inequality that includes $L$, i.e., to consider weak edges in the sparse graph.

The results given below were obtained for the sparse graph, $H = 0.7$, $p = 2$, and $L = 0$, were used as the values of the parameter, and the same result is obtained in the diapason of values $0.6 \le H \le 0.7$, $1 \le p$, and $L \le 0.05$. With regard to the effect of values of the parameters, at $p < 1$, clusters with the maximal size are destroyed; at unrestricted calculation times, higher values of $p$ may be used, even $p = +\infty$, i.e., conservation condition 1 may be discarded. When $L$ value exceeds 0.05, the number of edges starts to decrease quickly, the number of connected components to increase rapidly, and weak clusters are destroyed. At an insignificant increase in the value of $L$, the changes in clusters have a positive character. At $H \le 0.55$, some clusters are united and, at $H \ge 0.75$, some are destroyed.

It should be noted that the clustering algorithm may produce several variants; upon the construction of covering graph $D$, the algorithm produces several such graphs; and, at each increase in tree $D_{i,j}$ depth, the algorithm produces several variants of a new pair of trees. The number of these variants is small for the trees we considered (see the next paragraph). This allowed us to use the algorithm to select only one of the variants each time. At the same time, obstacles may arise that are overcome by hand by means of additional biological information. For example, the cluster of the L-subunit of protochlorophyllide reductase Ch1L was clustered by hand from a large cluster formed by the algorithm and combining various proteins that do not occur together in one plastome. The

**Table 4.** Graph *G* corresponds to the values in bold

| *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 | | | | | | | | | |

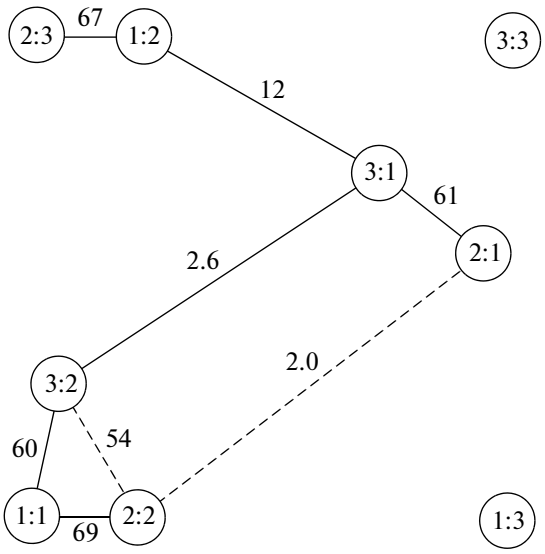


**Fig. 1.** Graph *G*. Length of the edge is approximately inversely proportional to its weight, i.e., more short edges correspond to more similarity between the corresponding proteins. Isolated vertices are placed randomly.

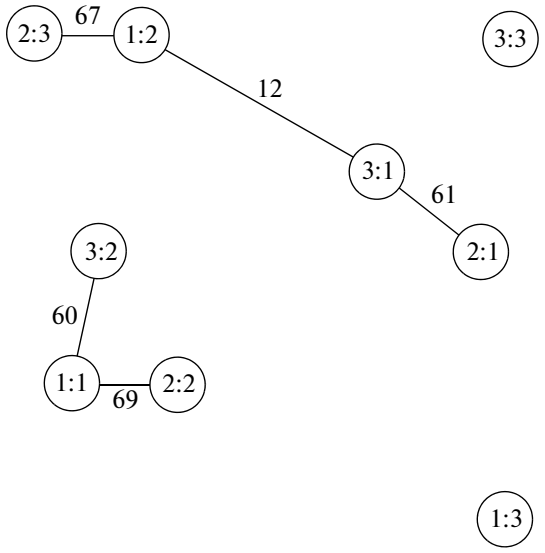


**Fig. 2.** Two trees after first division.

evolution of the *chlB*, *chlL*, and *chlN* genes that encode subunits of light-independent protochlorophyllide reductase was previously described [22]. Two more clusters were obtained by hand, one of which is made of fragments of the β"-subunit of bacterial-type RNA polymerase in Piroplasmida (*Babesia bovis* and *Theileria parva*), while the other is made of a kinase of algae *Rhodomonas salina* and *Heterosigma akashiwo*.

When constructing the tree *D*, edges with weight *s* may be considered competition that do not fall into *D*, while some other edge with this weight does. There are several hundred of these cases, while the total number of edges is on the order of 100000 in the sparse graph. The procedure of tree clustering is unambiguous for the discussed dataset. Generally speaking, taking into account the alternative variants will reduce manual calculations and may be useful.

Based on the clustering algorithm, a database was created to search for a protein based on its phylogenetic profile and other investigations of plastoms (available by address in [19]). The results obtained using this database and this algorithm were reported at the 53rd and 54th Scientific Conferences of the Moscow Institute of Physics and Technology and at the 50-Year Anniversary Institute for Information Transmission Problems Conference [8, 23, 24].

In order to verify our results and the construction of phylogenetic trees when investigating RNA polymerase, the MEGA 5 program package was used [25]. The search of the RNA polymerase subunit was performed using the BLAST program [26]; below, the corresponding expected value (e-value) will be designated as E.

The site of the reference [19] provides at least two functions, i.e., the search for a protein based on its phylogenetic profile, and a search based on a fragment of the amino acid sequence of all proteins encoded in plastids of the Rhodophyta branch that also contain a given fragment and clusters of these proteins. Instructions for use and examples of calculations are also given at this site.

## MATERIALS AND METHODS

Plastomes indicated in Table 1 were retrieved from the NCBI database; they include plastomes of recently sequenced diatomaceous algae [27, 28]. Some fragments of nuclear genes of *Eimeria tenella* and *Neospora caninum* Liverpool were obtained from the Sanger Institute database [29].

## RESULTS AND DISCUSSION

### Clustering of Plastome Proteins

We have considered numerous taxonomical groups of the Rhodophyta branch, which include all species of this branch represented in the GenBank NCBI database (Table 1). We considered 3426 proteins, of which 260 clusters were constructed with strictly more than one protein and 143 clusters were constructed with one-element clusters. The latter contain a total of 4% of all proteins; 11 clusters are entirely made from paralogous proteins to each other. Paralogs were not found in the majority of clusters, i.e., 359 clusters do not contain paralogs and 44 clusters do. The distribution of clusters by the number of species included is shown in Fig. 4.

We succeeded in clustering proteins that characterize several taxonomical groups, i.e., they are encoded

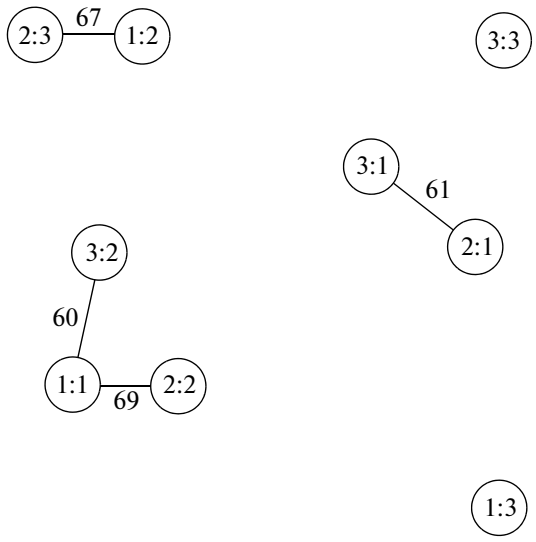


**Fig. 3.** A set of trees after the second division step.

in their plastomes and in no others. These proteins and the groups are listed below. Proteins common for plastomes of all the considered species comprise eight clusters, i.e., ribosomal proteins S2, S12, L2, L6, L14, and L16; elongation factor Tu; and the β-subunit of bacterial-type RNA polymerase. Ribosomal protein S19 has been determined in all considered species except Apicomplexa *Babesia bovis*.

Proteins encoded in plastids of Rhodophyta (*Cyanidioschyzon merolae*, *Cyanidium caldarium*, *Gracilaria tenuistipitata*, *Porphyra purpurea*, and *P. yezoensis*), but not in the other considered plastomes, form 24 clusters, including the third translation initiation factor; α-, β-, $\beta_{18}$-, γ-subunits of allophycocyanin; α- and β-subunits of phycocyanin; two structural proteins of phycobilisomes and the Ycf18 protein associated with degradation of phycobilisomes; thioredoxin; proteins of the acetyl-CoA carboxylase complex; prenyltransferase; acetyl glutamate kinase; ferrodoxin-dependent glutamate synthase; α-, β-subunits of pyruvate dehydrogenase E1; subunits of anthranylate synthase; α-subunit of tryptophan synthase; and hypothetical conservative proteins.

No protein has been found that would be specific to cryptophyte algae *Cryptomonas paramecium*, *Guillardia theta*, and *Rhodomonas salina*, as well as specific to Chromerida (*Alveolata* sp. CCMP3155 and *Chromera velia*).

Proteins specific for Apicomplexa of the group Piroplasmida (*Babesia bovis*, *Theileria parva*) comprise five clusters: two of which are weakly homologous of ribosomal proteins, while the other two are molecular chaperone homologs to ClpC (YP_00290851.1, XP_762692.1, YP_002290850.1, and XP_762693.1) and fragments of the β"-subunit of bacterial-type RNA polymerase (YP_002290845.1, XP_762712.1).

The group diatoms and dinotoms consists of *Durinskia baltica*, *Kryptoperidinium foliaceum*, *Fistulifera* sp. JPCC DA0580, *Odontella sinensis*, *Phaeodactylum tricornutum*, *Thalassiosira oceanica*, and *Thalassiosira pseudonana*. These include five clusters of diatomaceous algae, i.e., *Fistulifera* sp. JPCC DA0580, *P. tri-*

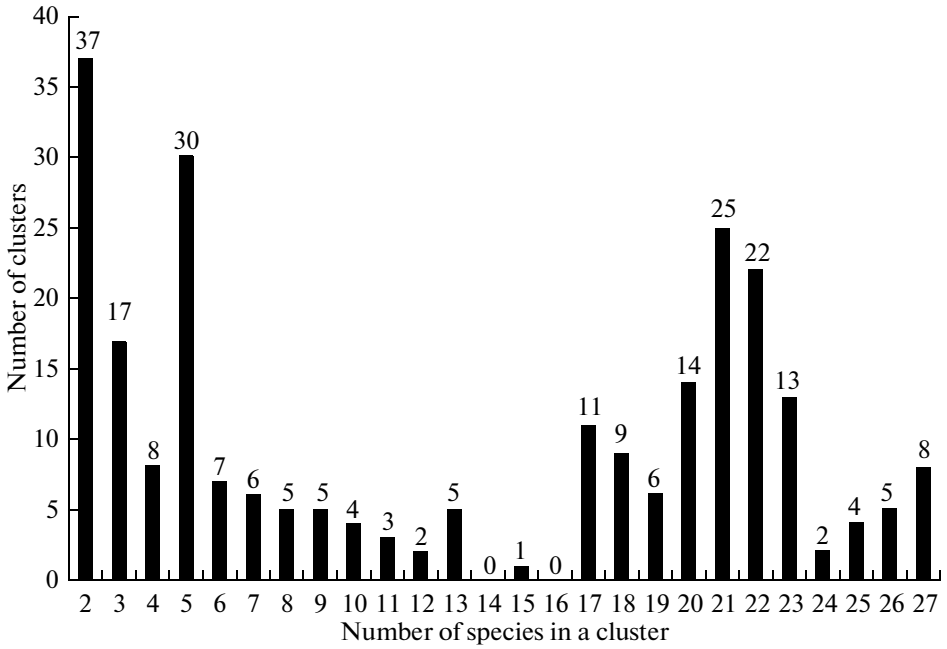


**Fig. 4.** Distribution of clusters by number of species. Number of plastom protein clusters is shown depending on the number of species represented in the cluster of the Rhodophyta branch; 154 clusters contain proteins only for one species.

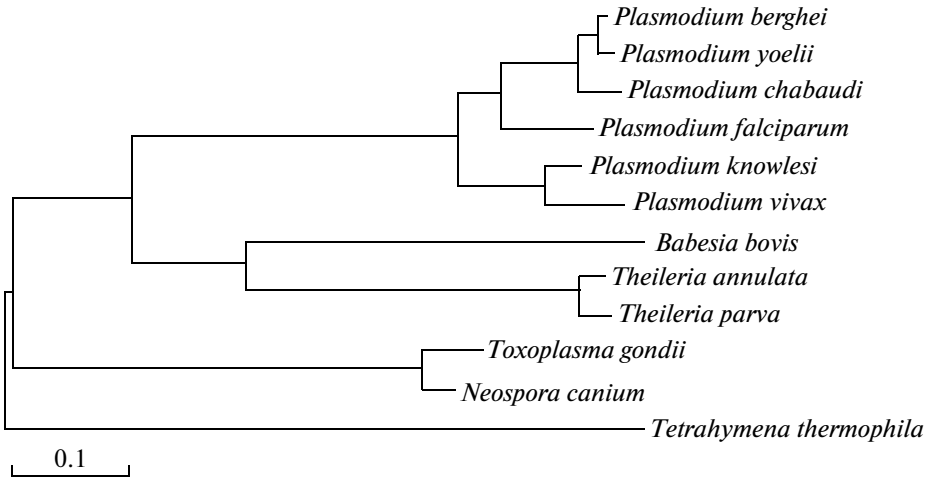


**Fig. 5.** Phage-type RNA polymerase tree in protozoan of Alveolata supertype.

*cornutum*, *O. sinensis, T. oceanica*, and *T. pseudonana*. Two clusters of the proteins are specific to this group; one cluster contains homologs of the Ycf88 protein, while the other is represented by pairs of paralogs that are homologous to the Ycf89 protein.

### Search for RNA Polymerase in the Nuclear Genomes of Apicomplexa

In strains *Toxoplasma gondii* ME49 (XP_002367014.1), *T. gondii* VEG (EEE31947.1), *T. gondii* GT1 (EEE23737.1), and in *Neospora caninum* (CBZ55882.1), one copy of phage-type RNA polymerase was found with the numbers indicated in the brackets. Proteins of the *T. gondii* strain ME49 and VEG coincide, and protein of the GT1 strain contains substitutions of amino acid residues in several positions and an insertion at the position of 347–354. We failed to determine phage-type RNA polymerase in *Eimeria tenella*.

Homologs of phage-type RNA polymerase are found in many Apicomplexa that do not belong to coccidia, including *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), and *Th. parva* (XP_766496.1). The tree of phage-type RNA polymerases is shown in Fig. 5. Nevertheless, no orthologous protein has been found for coccidia *Cryptosporidium parvum*, which, unlike many Apicomplexa, has no plastids.

Only one gene has been revealed in the nuclear genome of *Toxoplasma gondii* which encodes σ-subunit of the RNA polymerase. Its length is 1002 amino acids in strains ME49 and GT1, and 1001 amino acids in the VEG strain. The protein XP_002367841.1 strain ME49 is considered below. In the nuclear genome of *Neospora caninum*, the gene CBZ51366.1 codes for the σ-subunit of RNA polymerase with a length of 1206 amino acid residues. C-terminals of σ-subunit of RNA polymerase in *T. gondii* and *N. caninum* are extremely similar to each other; nevertheless, they do not have a significant similarity to σ-subunits of diatomaceous algae *Phaeodactylum tricornutum* CCAP 1055/1 and *Thalassiosira pseudonana* CCMP1335, golden algae *Aureococcus anophagefferens*, cryptyphyte algae *Guillardia theta*, and *Hemiselmis andersenii*. In coccidias, the σ-subunits closest to these σ-subunits are found in cyanobacteria *Cyanothece* sp. PCC 7822 (YP_003885480.1), *Microcoleus chthonoplastes* PCC 7420 (ZP_05024793.1), *Acaryochloris marina* MBIC11017 (YP_001519047.1), and in δ-proteobacterium *Desulfarculus baarsii* DSM 2075 (YP_003809216.1). Bacterial orthologs have lengths of 260–363 amino acid residues. The C-terminals of region 2, and the entire region 3 and N-terminals of the region 4 σ-subunit of RNA polymerase are well aligned in all species. Region 4 of *T. gondii, N. caninum* and *D. baarsii* was aligned along the entire length.

The orthologs of the σ-subunit of RNA polymerase were also found in protozoans of the order Haemosporida, including *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), and *Pl. chabaudi* (XP_739944.1). Other σ-subunits are not found in any of them. We failed to determine the σ-subunits of RNA polymerase for species of the order Piroplasmids, including *Theileria parva*, *Th. annulata*, and *Babesia bovis*. The tree of σ-subunits is shown in Fig. 6.

The peculiarity of Apicomplexa plastomes is the absence of α-subunits of bacterial-type RNA polymerases. We considered three species of coccidia, i.e., *Eimeria tenella*, *Toxoplasma gondii*, and *Neospora*

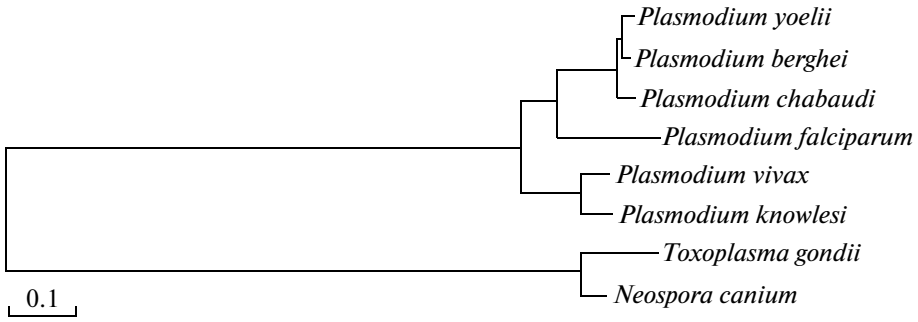


**Fig. 6.** Tree of σ-subunit of RNA polymerase in Apicomplexa.

*caninum*. Data on *T. gondii* and the discussed algae are available in the NCBI database. In *T. gondii* ME49 α-subunit is encoded in the nucleus, and the corresponding protein XP_002367289.1 has 836 amino acids residues. This protein differs by one position in stains of *T. gondii* ME49 and GT1. A related α-subunit ($E = 1.1 \times 10^{-71}$) was revealed in the nuclear genome of *E. tenella*, for which fragments of four exons on dev_EIMER_contig_00028796 contig were determined with coordinates 5283..5453..5682..6167, 6576..6785 and 7273..7965, respectively. A related α-subunit ($E = 9.9 \times 10^{-288}$) has been revealed in the nucleus genome of *N. caninum* with two exons on the Contig892 with coordinates 45655..47412 and 47940..48611, respectively.

As indicated above, the majority of conservative proteins are responsible for translation.

The NP_045121.1 protein in *Cyanidium caldarium* is included in the cluster that contains proteins YP_537023.1 of *Porphyra yezoensis* and NP_053952.1 of *P. purpurea*. These proteins have a relatively short conservative domain characteristic of the NtcA (Ycf28) transcriptional factor. The protein NP_849012.1 of *Cyanidioschyzon merolae* is a homolog of NtcA; nevertheless, it was not included into the NtcA cluster due to considerable differences, including the more conservative domain of the factor. There is even less similarity in the corresponding domain of NtcA and its homolog in *Gracilaris tenuistipitata*. This evolutionary change is associated with the elimination of the *glnB* gene from the plastome, the transcription of which is regulated by the NtcA factor in Rhodophyta *Porphyra* ssp. and *Cyanidium caldarium* [30].

The plastome of *Gracilaria tenuistipitata* contains genes *leuC* and *leuD* that encode large (YP_063540.1) and small (YP_063541.1) subunits of 3-isopropylmalat dehydrogenase, which are not found in the other considered plastomes. As noted above [31], this indicates the early division of taxonomic groups of Florideophyceae (which includes *G. tenuistipitata*) and Bangiophyceae in the order Rhodophyta.

The peculiarity of Apicomplexa plastomes is the absence of α-subunits of bacterial-type RNA polymerases; nevertheless, for these plastomes, their homologs were found in the nuclear genomes of the majority of Apicomplexa.

The existence of common proteins in diatomaceous algae and their closely related endosymbionts that are absent in plastids of other species allows one to assume that diatomaceous algae separated from other representatives of the Rhodophyta branch earlier than other species.

The nonconservativity of the majority of bacterial-type RNA polymerase subunits in Piroplasmida allows us to doubt the functional ability of this enzyme, since we have failed to find α-subunit in their nuclear genomes. In Piroplasmida, the transcription of the entire plastome is likely to be carried out by phage-type RNA polymerases. This indicates that the application of antibiotics that inhibit bacterial-type RNA polymerase in order to cope with Piroplasmids is ineffective. On the contrary, these antibiotics may be used against *Plasmodium* spp., *Toxoplasma gondii* and *Neospora caninum*.

The tree of σ-subunits of bacterial-type RNA polymerases in Apicomplexa, with the exception of species from Piroplasmida, well agrees with the tree of the species and the tree of phage-type RNA polymerases. The fact that Apicomplexa possess no more than one σ-subunit of RNA polymerase indicates the insignificant role of plastome regulation at the transcriptional level. It is likely that the regulation at the translational level or at the level of processing is more essential than has been proved by other studies [7].

Phage-type RNA polymerases are well aligned between each other in species of the genus *Plasmodium* forming a clade in protein tree; these polymerases also form separate Piroplasmida and Coccidia clades. Nevertheless, Coccidia phage-type RNA polymerases considerably differ from orthologous proteins of other Apicomplexa. On the contrary, phage-type RNA polymerase in Coccidia are close to orthologous proteins of tetrahymena, which has no plastids. It is possible to suggest that, in coccidia, phage-type RNA polymerases do not play a role in plastome transcription. We did not reveal any significant diversity of phage-type RNA polymerases in protozoans. Phage-type RNA polymerases in Apicomplexa are likely to be of

ancient origin and are not associated with plastid acquirement. On the contrary, a higher diversity of phage-type RNA polymerases is found in higher plants that work in various organelles [8, 11].

Similar results were obtained when studying the Chlorophyta branch of plants and algae, which will be published.

## ACKNOWLEDGMENTS

The study was supported by State Contracts of Ministry of Education and Science, Russian Federation (14.740.11.0624, 14.740.11.1053, 14.740.12.0830).

## REFERENCES

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 Biol*. **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. Balashov Yu.S. 1998. *Iksodovye kleshchi: parazity i perenoschiki infektsii* (Ixodid Ticks: Parasites and Infection Vectors). St. Petersburg: Nauka.
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. London: Ser. B*. **358**, 155–164.
6. Zhu G., Marchewka M.J., Keithly J.S. 2000. *Cryptosporidium parvum* appears to lack a plastid genome. *Microbiology*. **146**, 315–321.
7. Sadovskaya T.A., Seliverstov A.V. 2009. Analysis of the 5'-leader regions of several plastid genes in protozoa of the phylum Apicomplexa and red algae. *Mol. Biol*. (Moscow). **43**, 552–556.
8. Seliverstov A.V., Lyubetsky V.A. 2011. Evolution of RNA-polymerases and their promoters in plastids. *50-Year IITP Anniversary Conference, Moscow, Russia, September 15, 2011*. Moscow, pp. 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., and 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. http://orthomcl.cbil.upenn.edu/
14. http://roundup.hms.harvard.edu/browse/
15. http://www.omabrowser.org/
16. http://eggnog.embl.de/
17. http://inparanoid.sbc.su.se/
18. http://www.ncbi.nlm.nih.gov/COG/
19. 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. 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. *Eukar. Cell*. **7**, 550–553.
23. Zverkov O.A., Seliverstov A.V., Lyubetsky V.A. On an algorithm of protein clustering. *Trudy 53-ei nauchnoi konferentsii MFTI*. Proc. 53d Conf. Moscow Inst. of Physics and Technology (Moscow, 2010). Moscow: MFTI, 2010, part 1, vol. 1, pp. 118–119.
24. Zverkov O.A., Gorbunov K.Yu., Seliverstov, A.V. Lyubetsky V.A. Protein clustering with accounting for domen architecture. *Trudy 54-oi nauchnoi konferentsii MFTI*. Proc. 54th Conf. Moscow Inst. of Physics and Technology (Moscow, 2011), Section of Management and Applied Mathematics. Moscow: MFTI, 2011, vol. 2, pp. 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. 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. *Photosynth. Res*. **109**, 223–229.
29. http://www.sanger.ac.uk/
30. Lopatovskaya K.V., Seliverstov A.V., Lyubetsky V.A. 2011. NtcA and NtcB regulons in cyanobacteria and Rhodophyta chloroplasts. *Mol. Biol*. (Moscow). **45**, 522–526.
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.