Evolution of Chromosome Structures

R.A. Gershgorin, K.Yu. Gorbunov, A.V. Seliverstov, V.A. Lyubetsky

Institute for Information Transmission Problems of the Russian Academy of Sciences
(Kharkevich Institute)
gershgorin@iitp.ru

Abstract. An effective algorithm to reconstruct chromosomal structures is developed together with its computer implementation. The algorithm is applied to study chromosomal evolution in plastids of the rhodophytic branch and mitochondria of apicomplexan parasites.
The chromosomal structure is understood as an arbitrary set of linear and circular chromosomes where each gene is defined by the tail and head; the gene length, nucleotide composition, and intergenic chromosomal regions are not taken into account. We complement the standard operations with the operations of deletion and insertion of a chromosome fragment. The distance between chromosome structures is defined as the minimum total weight of the sequence of operations that transforms one structure into another where operation weights are not necessarily equal; and this sequence is called the shortest. Gene composition is variable, operation weights can be arbitrary and any paralogs are permissible.
By our algorithm we solve the following three tasks: (1) finding the distance and the corresponding shortest sequence; (2) finding the matrix of pairwise distances between structures from a given set, and generating the optimal evolutionary tree for the matrix; (3) reconstructing the ancestral structures based on the structures at the tree leaves.

Keywords: evolution of chromosome structures, rhodophytic branch plastids, apicomplexan mitochondria, model of chromosome structures

1 Introduction

Rapidly increasing number of sequenced plastid genomes makes it possible to advance our knowledge about the evolution of both algae and plastid-bearing non-photosynthetic protists. Apicomplexan plastids are included in the rhodophytic branch [1]. The latter include agents of serious protozoan infections, in particular, malaria and toxoplasmosis. For instance, *Toxoplasma gondii* is a medically and veterinary important apicomplexan with an extremely wide distribution [2, 3]. Plastids serve as the target for efficient therapeutic treatment, fast generation of non-virulent strains for vaccine production, etc. This equally applies to the studies on mitochondrial evolution, particularly, in sporozoans, which include the agents of malaria, babesiosis, and theileriosis. Unfortunately, in the latter case the theoretical analysis is substantially limited by the low number of species with sequenced mitochondrial genomes, which largely belong to the class Aconoidasida.

David Sankoff *et al* were the first to apply the distance-based method to reconstruct phylogeny from gene orders [4]. In their paper, the phylogenetic tree of animals was generated using the order of genes in mitochondrial genomes. Then mitochondrial gene orders were reconstructed at the ancestral tree nodes with the minimum total number of breakpoints for all edges, which is a marker of tree quality. Recall that this number is the number of pairs of gene ends adjacent in one structure and not adjacent in the other. The tree and the ancestral orders reconstruction were generated by exhaustive search and traveling salesman heuristic, respectively; 11 species were considered.

Since then, many studies compared chromosome structures in the context of their evolution. An extensive series of works by Pavel Pevzner and his school, in which the algorithms of chromosome rearrangements included inversions as well as some transversions and translocations, are important (see chapter 4 in [5] for review).

The detection of syntenic blocks and the construction of a phylogenetic tree from them were considered elsewhere [6]. However, this approach involved no distances between chromosome structures and the reconstruction of ancestral chromosome structures was not considered. Shao *et al* discussed the standard chromosome transformations with an extra duplication operation, which closes the copy of a chromosome fragment into a circle or inserts the copy into another chromosome location [7]. It is shown that their algorithm finds a sequence with the minimum number of operations in the absence of the extra operation. Another publication [8] addresses the problem of finding the minimum number of standard operations to transform one chromosome structure into another for an invariable set of genes and without deletions or insertions of genes but with paralogs. Namely, both structures have the same number of paralogs for each gene, and bijections between the corresponding pairs of paralog sets are searched. A way to reduce this task to the task of integer linear programming, which still remains NP-hard, is proposed. Wang *et al* [9] consider constructing a phylogenetic tree from a matrix of pairwise distances between chromosome structures in the leaves. The structures include one circular chromosome each, and all structures have an invariable set of genes without paralogs. Two definitions of the distance between structures are considered: the minimum number of inversion operations to transform one structure into another and the number of breakpoints. Corrections based on statistical evaluations are proposed for both definitions, which improve fast methods for tree construction (such as neighbor joining or UPGMA). The approach is also applicable for a single linear chromosome. Wang *et al* [10] compared three definitions of the distance between structures: invertional, breakpoint-based, and statistical. The latter is proposed as the optimal approach. It relies on the statistical adjustment of the inversion distance and was also described in [11]. Operation weights are not considered in any of these studies or, in other words, all weights are supposed to be equal.

Baudet *et al* [12] analyzed the transformation of one circular chromosome (without paralogs) into another using the inversion operation. A heuristic generating the shortest sequence of inversions was proposed; an operation weight was

used that depends on the inversion length and symmetry, which is the equidistance of inversion ends from the origin of replication.

2 Materials and Methods

All data were obtained from GenBank. In the analysis of the evolution of plastid chromosome structure, the genes available in many species and encoding proteins with a certain function were used: chaperone *clpC*; subunits of photosystem I *psaA*, *psaB*, *psaC*, *psaD*, *psaE*, *psaF*, *psaI*, *psaJ*, *psaK*, *psaL*, and *psaM*; subunits of photosystem II *psb28*, *psb30*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbN*, *psbT*, *psbV*, *psbX*, *psbY*, and *psbZ*; rubisco large subunit *rbcL*; RNA polymerase subunits *rpoA*, *rpoB*, *rpoC1*, *rpoC2*, and *rpoZ*; ribosomal proteins *rpl1*, *rpl2*, *rpl3*, *rpl4*, *rpl5*, *rpl6*, *rpl9*, *rpl11*, *rpl12*, *rpl13*, *rpl14*, *rpl16*, *rpl18*, *rpl19*, *rpl20*, *rpl21*, *rpl22*, *rpl23*, *rpl24*, *rpl27*, *rpl28*, *rpl29*, *rpl31*, *rpl32*, *rpl33*, *rpl34*, *rpl35*, *rpl36*, *rps1*, *rps2*, *rps20*, *rps3*, *rps4*, *rps5*, *rps6*, *rps7*, *rps8*, *rps9*, *rps10*, *rps11*, *rps12*, *rps13*, *rps14*, *rps16*, *rps17*, *rps18*, and *rps19*; and elongation factor *tufA*. Paralogs of the *psbY* gene can be found in *Odontella sinensis*, *Phaeodactylum tricornutum*, *Thalassiosira pseudonana*, *Thalassiosira oceanica*, *Ulnaria acus*, *Asterionella formosa*, *Asterionellopsis glacialis*, *Didymosphenia geminata*, *Lithodesmium undulatum*, *Eunotia naegelii*, *Chaetoceros simplex*, *Roundia cardiophora*, *Cerataulina daemon*, and *Thalassiosira weissflogii*. Paralogs of the *clpC* gene can be found in *Theileria parva*, *Babesia bovis*, *Chromera velia*, *Thalassiosira oceanica*, *Nannochloropsis gaditana*, *Nannochloropsis granulata*, *Nannochloropsis oculata*, *Nannochloropsis salina*, *Nannochloropsis limnetica*, *Nannochloropsis oceanica*, and *Rhizosolenia imbricata*. Consecutive paralogs of the *rpoC2* can be found in *Theileria parva*, *Leucocytozoon caulleryi*, and *Plasmodium chabaudi*. In *Rhizosolenia imbricata*, a long duplication includes the *psbA*, *psaC*, *rps6*, *clpC*, *rps10*, *rps7*, and *rps12* genes. Notice formal errors in gene names in GenBank annotations: *rpo* instead of *pro1* in *Nannochloropsis gaditana*, *rpoC* instead of *pro1* in *Cyanidioschyzon merolae*, and *rpo2-n-terminal* instead of *pro2* in *Babesia bovis*.

The evolution of mitochondrial chromosome structure was studied in the sporozoan class Aconoidasida composed of subclasses Haemosporida and Piroplasmida (Table 1). Here even closely related species can have linear and circular chromosomes (Table 1, column 3). No paralogs were found in the mitochondria under consideration.

Proteins were clustered using the algorithm described elsewhere [13, 14] with the parameters $E = 0.001$, $L = 0$, and $H = 0.6$. Orthology and paralogy of genes were determined from thus obtained clustering.

The chromosomal structure is understood as an arbitrary set of linear and circular chromosomes where each gene is defined by the head and tail; the gene length, nucleotide composition, and intergenic chromosome regions are not taken into account [15]. We complement the standard operations with the operations of *deletion* and *insertion* of a chromosome fragment. The distance between chromosome structures is defined according to the parsimony principle as the minimum

total weight of the sequence of operations that transforms one structure into another where operation weights are not necessarily equal; and the sequence is called the *shortest*. Specifically, the *biological distance* between structures here is the arithmetic mean of the total weights of the shortest sequences of operations transforming one structure into another because the total weights can be unequal. Gene composition is variable, operation weights can be arbitrary and any paralogs are permissible.

For the case of variant gene composition the *breakpoint distance* definition should be improved. The breakpoint distance between two structures with a numbering of paralogs is the number of breakpoints plus the number of genes, which are present at one structure and absent at another one. Notice that the breakpoint distance can be effective to help compute the biological distance and chromosomal reconstruction.

The computation of the biological distance and the corresponding shortest sequence is the *first task*. Hereafter, the biological distance is implied unless the breakpoint distance is explicitly mentioned.

The *second task* is to compute the matrix of pairwise distances between structures from a given set, and to generate the optimal evolutionary tree for the matrix. The biological distance here is the minimum biological distance for all numberings. The *third task* is to reconstruct the ancestral structures based on the structures at the tree leaves. Thus, given a tree (not necessarily binary) with individual chromosome structures specified in each leaf; gene composition is arbitrary and paralogs are allowed. The structures with paralog numberings at the tree nodes should be reconstructed to minimize the function: total distance between the structures with paralog numberings at the edge termini (for all edges). Internal node structures can only include genes represented in the leaves. Firstly we solve the third task for the breakpoint distance between structures at the edge termini. The subsequent section presents the solving algorithm. Then the reconstruction algorithm described elsewhere ([15], Biological Distance section) is applied to the resulting breakpoint solution with biological distance between structures at the edge termini.

For the three tasks exact effective solving algorithms are developed and applied to obtain the presented below results. Their computer implementations are available at http://lab6.iitp.ru/en/chromo/.

2.1 Chromosome reconstruction with the breakpoint distance

The node to the root defines an *edge-tail*; an *edge-head* is defined correspondingly. The set of all orthologs of gene k in all leaves of the evolutionary tree combined with the set of their paralogs in all leaves is called the *paralogous group* of k. Elements of this group are called (here, for convenience) *k-paralogs*. These paralogs are numbered from $k.1$ to $k.n_k$. Each tree node corresponds to a structure that can include only k-paralogs for any k. Hereafter, the nodes and their structures are identified.

Paralogs numbering is defined for each edge and each gene k represented in both termini i and j of the edge as bijection f between two subsets of the sets of

all k-paralogs in i and in j. If $i = f(j)$ we assume that paralog $k.i$ at the edge-head descends from paralog $k.j$ at the tail of this edge, otherwise i appears de novo; paralogs in the complement of the domain of f were lost. Each paralog is assigned a name of the type $k.l$ where $1 \leqslant l \leqslant n_k$. If structures and numberings are fixed along the tree one can rename $i = f(j)$ to j at all edges such that all bijections become identical. Note that names at the root are assigned arbitrarily; gene names at the leaves are usually renumbered accordingly. Moreover, it is convenient to *formally* reckon that at each node all paralogs of each gene are considered. Therefore the task consists of both optimal coordination of paralogs by their numberings at all edges and determination of structures themselves.

In this section, the task is solved based on Boolean linear programming. In this context, let us introduce the variables, linear constraints, and linear function to be minimized. Let us introduce a variable z_{kije} (the first index will be omitted for brevity) for each edge e of the tree and each ordered pair $k.i$, $k.j$ of paralogs of a gene k; this variable equals 1 if $k.i = f(k.j)$ where $k.i$ is at the edge-head, $k.j$ is at the edge-tail, otherwise it equals 0. It is convenient to reckon that i and j formally range from 1 to n_k. Notice that the $k.i$-th or $k.j$-th paralogs can actually be missing in the corresponding structures. The correspondence between paralogs defined by z_{ije} should be bijective, which is expressed as linear equations: for fixed i and e, the sum over j of z_{ije} values equals 1, and this is also true for fixed j and e. The variable determines a paralogs numbering.

Let us introduce a variable $x_{kl\nu}$ for each node ν of the tree and each unordered pair k, l of different gene ends; this variable equals 1 if these ends are adjacent at the node ν, otherwise it equals 0. Each end can be adjacent to no more than one other end, which is expressed as linear inequalities: for fixed k and ν, the sum over l of $x_{kl\nu}$ values does not exceed 1, and this is also true for fixed l and ν. The variable determines the structure at a node, i.e. an arrangement of structures along the tree.

Let us introduce a variable $y_{k\nu}$ for each node ν of the tree and each gene k; this variable equals 1 if the gene k is absent from the structure of this node; otherwise it equals 0. The variables x_{kl} and y_k are given at the leaves. The ends of missing genes are not adjacent, which is expressed as linear inequalities $x_{ij\nu} \leqslant 1 - y_{k\nu}$, where i or j is an end of the gene k. For each gene k and each its end i, instead of these inequalities one inequality can be used $\sum\limits_{j \neq i} x_{ij\nu} \leqslant 1 - y_{k\nu}$ where j runs over all ends at an arbitrary node ν. The variable determines a source gene composition at all nodes.

Let us call two pairs of gene ends, k, l and m, n (belonging to the edge-head and edge-tail of an edge, respectively) *similar* if both pairs are two tails, or two heads, or tail and head of a gene (in the latter case, the ends must belong to the same or different genes in both pairs); in addition, genes i_1 and j_1 with the ends k and m, respectively, and genes i_2 and j_2 with the ends l and n, respectively, (or vice versa, the genes with the ends k and n and the genes with the ends l and m) should belong to the same paralogous group. Let us introduce a variable s_{klmne}, for each edge e and two similar pairs k, l and m, n of gene ends; this variable equals 1 if these ends belong to the

corresponding (as defined by the z-type variables) paralogs and the ends are adjacent at one terminus of the edge e and not adjacent at the other. The condition for s_{klmne} is expressed as linear inequalities; they are presented below for the case when the ends in both pairs are tails or heads of paralogs from the same paralogous group at the edge-head (other cases are analogous): if $e = (u, v)$ then $s_{klmne} \geqslant x_{kl\nu} - x_{mnu} - (1 - z_{i_1 j_1 e}) - (1 - z_{i_2 j_2 e})$, $s_{klmne} \geqslant x_{mnu} - x_{kl\nu} - (1 - z_{i_1 j_1 e}) - (1 - z_{i_2 j_2 e})$, $s_{klmne} \geqslant x_{kl\nu} - x_{mnu} - (1 - z_{i_1 j_2 e}) - (1 - z_{i_2 j_1 e})$, $s_{klmne} \geqslant x_{mnu} - x_{kl\nu} - (1 - z_{i_1 j_2 e}) - (1 - z_{i_2 j_1 e})$.

It is not followed from these inequalities that $s_{klmne} = 0$, thus s_{klmne} can take on the value of 0 or 1, although the value of 0 is desirable considering that the sum of these variables is minimized.

Let us introduce a variable s_{ije} for each edge $e = (u, v)$ and two genes i (at the edge-tail) and j (at the edge-head); this variable equals 1 if the genes correspond to each other according to the values of z-type variables and one of genes is present and the other gene absent. This condition is expressed as linear inequalities $s_{ije} \geqslant y_{i\nu} - y_{ju} - (1 - z_{ije})$ and $s_{ije} \geqslant y_{ju} - y_{i\nu} - (1 - z_{ije})$. Finally, we have to minimize the linear function equal to the sum of all *Boolean* variables s_{klmne} and s_{ije} for the linear constraints specified above. Any solution of this problem defines an arrangement with the minimum total weight of structures on the tree together with the numberings of all paralogs.

Thus, the task of this section was reduced to the minimization of a linear polynomial over the set of unit cube vertices.

The proposed reduction of chromosomal structure reconstruction to Boolean linear programming can be easily generalized for the case when the breakpoint weights (the transition along an edge to adjacent gene ends and vice versa as well as the transition to the gene loss and vice versa) are different. For that, each variable s_{klmne} should be replaced with two variables $s1_{klmne}$ and $s2_{klmne}$. The former will be limited to inequalities of the following type: $s1_{klmne} \geqslant x_{kl\nu} - x_{mnu} - \ldots$; the latter, to inequalities $s2_{klmne} \geqslant x_{mnu} - x_{kl\nu} - \ldots$. Similarly, variables s_{ije} should be replaced with $s1_{ije}$ and $s2_{ije}$. It allows to multiply the new variables to coefficients that reflect the event weights in the function to minimize.

The tree of plastids generated by our approach using the biological distance is shown in Fig. 1. Figs. 2 and 3 present the reconstruction of the same chromosome structures given in the leaves for two subtrees (named *small* and *large*) of this tree using the biological distance. The large tree was reconstructed using two millions of variables and four millions of linear equalities and inequalities.

3 Results

3.1 Evolution of chromosome structures in plastids of rhodophytic branch

The tree of plastids is in good agreement with previously published data, in particular, with the corresponding trees of species. The most significant distinctions are special tree positions of photosynthetic alveolate *Chromera velia*

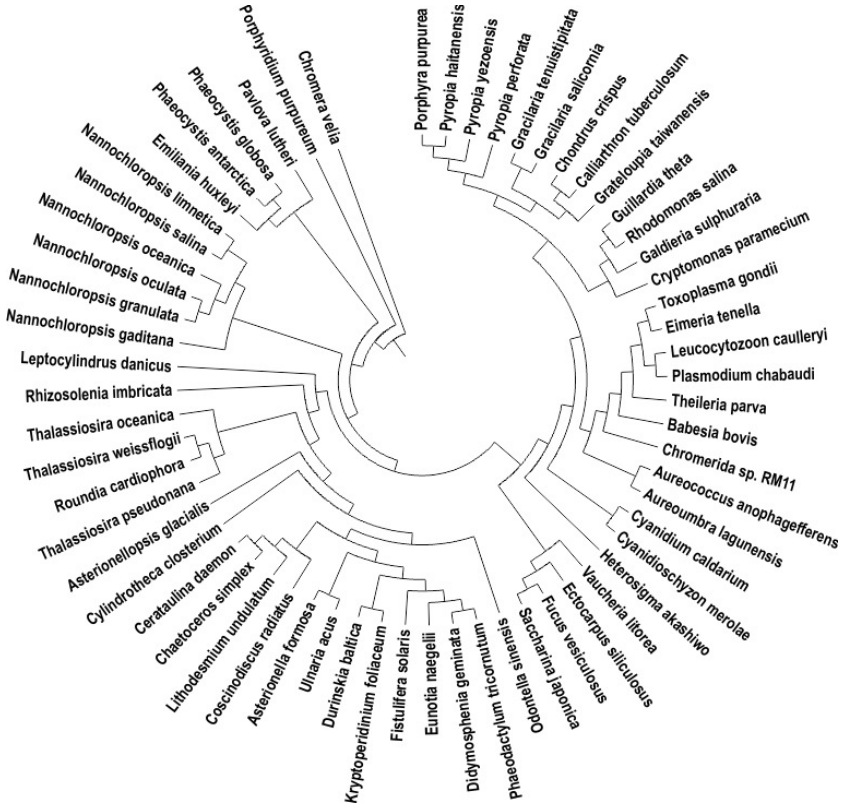

Fig. 1. Tree of chromosome structures of rhodophytic plastids generated by our algorithm using biological distance

and rhodophytic alga *Porphyridium purpureum*, whose order of genes substantially differs from that in related species. This has been mentioned previously in the study of the *moeB* gene regulation [16]. A separate clade was formed by plastids of haptophycean algae *Emiliania huxleyi*, *Pavlova lutheri*, *Phaeocystis antarctica*, and *P.globosa*. The third separate clade included the *Nannochloropsis genus*, which constitute an isolated portion of the large Stramenopiles phylum [17].

All diatoms composed a large clade also including certain stramenopiles as well as alveolate species *Durinskia baltica* and *Kryptoperidinium foliaceum*, whose plastids are of tertiary origin descending from diatom ones [18].

Another large clade was formed by plastids of rhodophytic algae excluding *Porphyridium purpureum*, cryptophytes, certain alveolates, and stramenopiles *Aureococcus anophagefferens* and *Aureoumbra lagunensis* [19]. This clade is closely related to other stramenopile algae: raphidophyte *Heterosigma akashiwo* [20], xanthophyte *Vaucheria litorea*, and brown algae *Ectocarpus siliculosus*, *Fucus vesiculosus*, and *Saccharina japonica* [21, 22]. The alignment of the upstream re-

gions of the *ftsH* gene in *H.akashiwo*, *V.litorea*, *E.siliculosus*, *F.vesiculosus*, and *S.japonica* agrees with the similar mutual arrangement of genes in the plastids of these species.

The alveolate species whose plastids are close to those of rhodophytic algae include all considered sporozoans as well as the photosynthetic alveolate *Chromerida* sp. RM11. The common origin of these plastids has been previously confirmed by protein alignment [23, 1]. The common origin of their plastids (excluding apicoplasts in piroplasmids *Babesia bovis* and *Theileria parva*) is further confirmed by the absence of the *ycf16* (*sufC*) gene located in the operon with the *ycf24* (*sufB*) gene in most rhodophytic plastids. In addition, we have predicted a uniform *ycf24* expression control in plastids of sporozoans and certain rhodophytic algae [24], which corroborates the close positions of these species on the generated tree.

A significant variation between plastids is observed among stramenopiles. The distinction of haptophytes agrees with the independent origin of plastids in Haptophyta and Stramenopiles proposed previously [25]. However, the independent origin of cryptophyte plastids is not confirmed. Overall, one can propose that plastids of rhodophytic branch are monophyletic and descend from plastids of rhodophytic algae, while this statement seems questionable for cryptophytes and sporozoans.

The tree of plastids based on the order of genes [26, Fig. 6b] represented only 7 rhodophytic species; it has a separate clade of rhodophytic algae with a subdivision of cyanidales, *Cyanidioschyzon merolae* and *Cyanidium caldarium*. This pattern was reproduced in our tree based on a larger volume of data. The mutual arrangement of rhodophytic algae and diatom *Odontella sinensis* also matches that in our tree. However, the tree in [26] differs from our tree by the position of haptophyte *Emiliania huxleyi* and cryptophyte *Guillardia theta* relative to rhodophytic algae and the diatom.

3.2 Reconstruction of chromosome structures in rhodophytic plastids along the tree of their evolution

For brevity, the reconstructions in two subtrees of the tree shown in Fig. 2 and 3 are presented: from the common ancestor of *Leptocylindrus danicus* and *Odontella sinensis* (hereafter, small tree) and from the common ancestor of *Porphyra purpurea* and *Saccharina japonica* (hereafter, large tree). The evolutionary scenario for the small tree is shown in Fig. 2. One can see that all chromosomes are circular.

The evolutionary scenario for the large tree is shown at Fig. 3. As previously, all chromosomes are circular; the ancestral structures include a single chromosome in the subtree up to the common ancestor of *Porphyra purpurea* and *Cryptomonas paramecium*; while most structures contain several chromosomes in the remaining part of the subtree. This can point to active chromosome rearrangements in ancestral species in this part of the subtree.

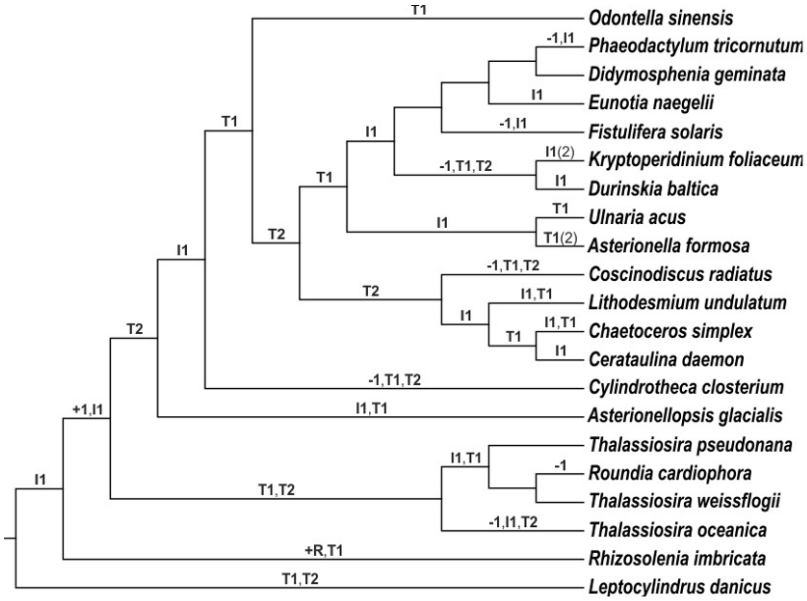

Fig. 2. Evolutionary scenario of chromosome structures along of the small tree. The following events are shown on edges: -1 – a loss of one of two paralogs of gene *psbY*; +1 – emergence of a paralog of gene *psbY*; +R – emergence of inverted repeat of a chromosome segment; I1 – inversion of a chromosome segment; T1 – transversion of a chromosome segment; T2 – translocation of a chromosome segment. In parentheses the number of the events is given when it is greater than 1

3.3 Evolution and reconstruction of mitochondrial structures in sporozoans

The tree of chromosome structures of mitochondria in sporozoan class Aconoidasida was generated by our method using the biological distance from data shown in Table 1. The tree shown in Fig. 4 consists of two clades including mitochondria of piroplasmids (genera *Babesia* and *Theileria*) and haemosporids (genera *Plasmodium* and *Leucocytozoon*), respectively. Two genera *Plasmodium* and *Leucocytozoon* cannot be resolved on the tree, in particular, due to the presence of linear and circular mitochondrial DNA in both of them.

The phylogenetic reconstruction of mitochondrial chromosome structures in Aconoidasida is shown in Table 2.

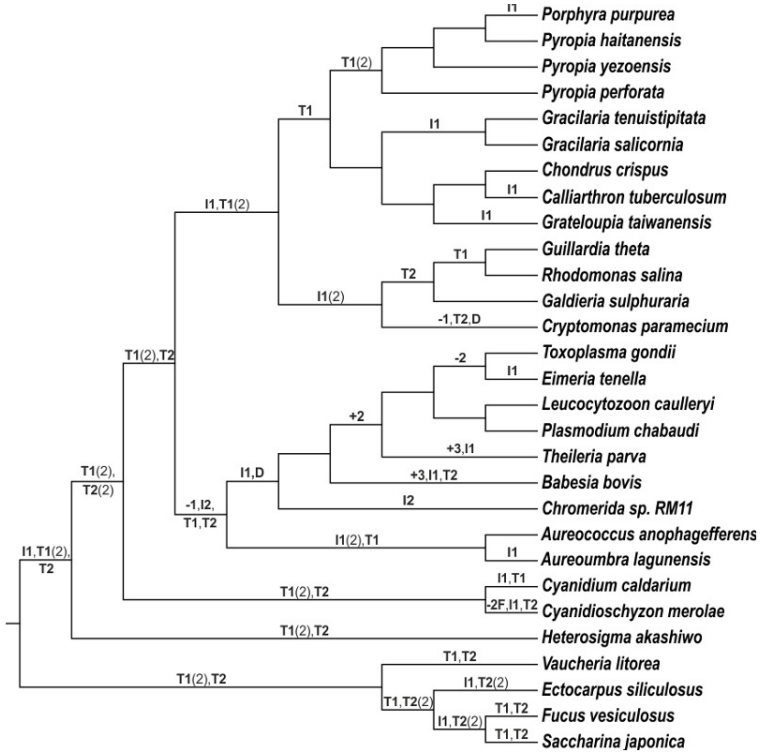

Fig. 3. Evolutionary scenario of chromosome structures along of the large tree. The following events are shown on edges: -1 – a loss of gene *psbY*; -2 – a loss of one of two paralogs of gene *rpoC2*; 2F – fusion of two paralogs of gene *rpoC2* into one large gene; +2 – emergence of a paralog of gene *rpoC2*; +3 – emergence of a paralog of gene *clpC*; I1 – inversion of a chromosome segment; T1 – transversion of a chromosome segment; T2 – translocation of a chromosome segment; I2 – insertion of a chromosome segment; D – disappearance of a chromosome segment. In parentheses the number of the events is given when it is greater than 1

Table 1: **Mitochondrial chromosome structures in the class Aconoidasida.** Circular and linear chromosomes are marked by C and L, respectively. In the list of genes, asterisk indicates the complementary strand relative to that specified in GenBank. Genome compositions were checked by BLAST alignments and Rfam database search. The rightmost column shows the gene order using standard gene names.

Haemosporida subclass		
Species	Locus in GenBank	Composition
Leucocytozoon fringillinarum	FJ168564.1	ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss6 ls7 ss4 (C)
Leucocytozoon majoris	FJ168563.1	ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss6 ls7 (C)

Leucocytozoon sabrazesi	NC_009336.1	ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Plasmodium berghei	NC_015303.1	ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Plasmodium falciparum	NC_002375.1	ss3 ls3 ls9 ss2 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)
Plasmodium floridense	NC_009961.2	ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)
Plasmodium fragile	AY722799.1	ls1 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb (C)
Plasmodium gallinaceum	NC_008288.1	ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Plasmodium juxtanucleare	NC_008279.1	ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Plasmodium knowlesi	NC_007232.1	ls1 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb (C)
Plasmodium mexicanum	NC_009960.2	ss3 ls3 ls9 ss2 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)
Plasmodium reichenowi	NC_002235.1	ss3 ls3 ls9 ss2 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)
Plasmodium relictum	NC_012426.1	ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (C)
Plasmodium simium	NC_007233.1	ls1 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 cox1 cytb ls8 ss5 ss1 (C)
Plasmodium vivax	NC_007243.1	ls1 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb (C)
Piroplasmida subclass		
Babesia bovis	NC_009902.1	cox1 *cox3 ls1 *ls2 *ls3 *cytb *ls4 ls5 (L)
Theileria parva	NC_011005.1	cox1 *cox3 ls1 *ls3 *cytb *ls5 ls4 (L)
Theileria annulata	CR940346.1	cox1 *cox3 ls1 *ls3 *ls2 *cytb *ls5 ls4 (L)

Table 2: **Phylogenetic reconstruction of mitochondrial chromosome structures in sporozoan class Aconoidasida.** Circular and linear chromosomes are marked by C and L, respectively. The upper line in each cell shows a non-leaf tree vertex by the first and the last leaves. The lower line shows the chromosome structure in the vertex (the order of rows corresponds to the traversal of the tree in Fig. 4). The leaves are labeled by (l), their chromosomal structures are feeded to the input of our algorithm.

Leucocytozoon majoris – Theileria parva
ls2 *ss2 *ls9 ss4 ss6 ls7 ls6 ss3 ls3 *ls1 cox3 *cox1 *ss1 *ss5 *ls8 *cytb *ls5 *ls4 (L)
Leucocytozoon majoris – Plasmodium berghei
ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (C) | ls2 (L)
Leucocytozoon majoris – Plasmodium simium
ss4 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 (C)
Leucocytozoon majoris – Plasmodium mexicanum
ss4 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 (C)
Leucocytozoon majoris – Leucocytozoon fringillinarum
ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss6 (C)
Leucocytozoon majoris – Plasmodium vivax
ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss6 (C)
Leucocytozoon majoris – Plasmodium knowlesi
ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss6 (C)
Leucocytozoon majoris (l)
ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss6 ls7 (C)
Plasmodium fragile (l)
ls1 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb (C)
Plasmodium knowlesi (l)
ss4 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 (C)
Plasmodium vivax (l)
ls1 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb (C)
Leucocytozoon fringillinarum (l)
ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss6 ls7 ss4 (C)
Plasmodium floridense – Plasmodium mexicanum
ss4 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 (C)
Plasmodium floridense – Plasmodium relictum
ss4 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 (C)
Plasmodium floridense (l)
ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)
Plasmodium relictum (l)
ss3 ls3 ls9 ss2 ls4 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (C)
Plasmodium reichenowi – Plasmodium mexicanum
ss3 ls3 ls9 ss2 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)
Plasmodium reichenowi (l)
ss3 ls3 ls9 ss2 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)
Plasmodium falciparum – Plasmodium mexicanum
ss3 ls3 ls9 ss2 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)
Plasmodium falciparum (l)
ss3 ls3 ls9 ss2 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)

Plasmodium mexicanum (*l*) ss3 ls3 ls9 ss2 *cox3 ls8 ss5 ss1 cox1 cytb ls1 ss4 ss6 ls7 (L)
Plasmodium simium (*l*) ls1 ss6 ls7 ss3 ls3 ls9 ss2 ls4 *cox3 cox1 cytb ls8 ss5 ss1 (C)
Plasmodium juxtanucleare – Plasmodium berghei ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Plasmodium juxtanucleare (*l*) ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Plasmodium gallinaceum – Plasmodium berghei ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Plasmodium gallinaceum (*l*) ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Leucocytozoon sabrazesi – Plasmodium berghei ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Leucocytozoon sabrazesi (*l*) ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Plasmodium berghei (*l*) ls1 ss4 ss6 ls7 ls6 ss3 ls3 ls9 ss2 ls4 ls5 *cox3 ls8 ss5 ss1 cox1 cytb ls2 (L)
Babesia bovis – Theileria parva ls4 ls5 cytb ls3 *ls1 cox3 *cox1 (L)
Babesia bovis (*l*) cox1 *cox3 ls1 *ls2 *ls3 *cytb *ls4 ls5 (L)
Theileria annulata – Theileria parva ls4 ls5 cytb ls3 *ls1 cox3 *cox1 (L)
Theileria annulata (*l*) cox1 *cox3 ls1 *ls3 *cytb *ls5 ls4 (L)
Theileria parva (*l*) cox1 *cox3 ls1 *ls3 *ls2 *cytb *ls5 ls4 (L)

4 Discussion

A high-level model of chromosome structure was proposed together with computer programs that allow its efficient utilization. A database of protein families encoded in rhodophytic plasmids was generated
(available at http://lab6.iitp.ru/ppc/redline63/). The scenarios of chromosome rearrangements were deduced in rhodophytic plastids and sporozoan mitochondria. The scenarios, in particular, demonstrate the similarity of chromosome structures in sporozoan apicoplasts and rhodophytic plastids, which agrees with the previously proposed common origin of expression regulation in a few genes from these species, including the common pattern of translation initiation regulation for genes coding for DNA-directed RNA polymerase beta chain and the protein SufB involved in iron-sulfur cluster formation. The similarity of chromosome structures is observed in rhodophytic and cryptophytic plastids. On the other hand, our results indicate an early and independent segregation of diatom and haptophyte plastids.

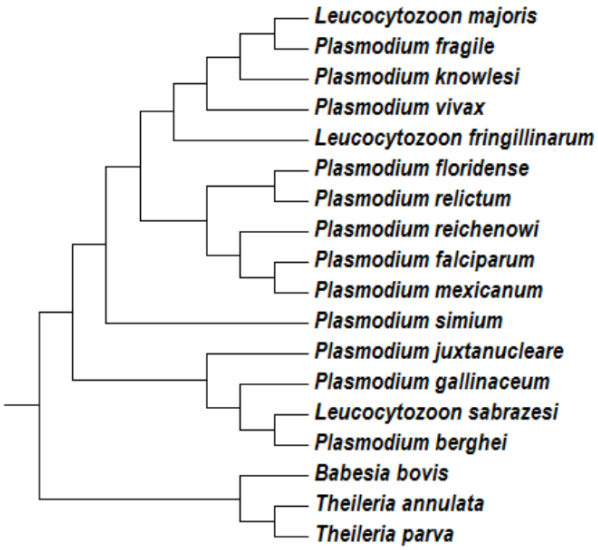

Fig. 4. The tree of chromosome structures of mitochondria in sporozoan class Aconoidasida

Chromosome structures in plastids of the rhodophyte alga *Porphyridium purpureum* and the photosynthetic alveolate alga *Chromera velia* deviate considerably from those in their relatives. In such cases chromosomes cannot be used to infer phylogenetic relationships but can provide comparative information for understanding the role of chromosome rearrangement in gene expression regulation. Such analysis was published for plastids of higher plants [27].

Chromosome rearrangements can considerably affect patterns of gene expression, particularly due to RNA polymerases competition [28]. In chromosomes with labile structures transcription terminators are naturally expected to occur in between genes, and gene expression be largely regulated at the translation level, which was described for many plastids [29].

The model and computer programs are applicable to study other genomes as a method to explore the evolution of chromosome structures.

Acknowledgements

The study was supported by the Russian Scientific Fund (14-50-00150).

References

1. Janouškovec, J., Liu, S.L., Martone, P.T., Carre, W., Leblanc, C., Collen, J., Keeling, P.J.: Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. PLoS ONE. 8(3), E59001 (2013)

2. Andenmatten, N., Egarter, S., Jackson, A.J., Jullien, N., Herman, J.P., Meissner, M.: Conditional genome engineering in Toxoplasma gondii uncovers alternative invasion mechanisms. Nat Methods. 10, 125–127 (2013)
3. Ermak, T.N., Peregudova, A.B., Shakhgil'dian, V.I., Goncharov, D.B.: Cerebral toxoplasmosis in the pattern of secondary CNS involvements in HIV-infected patients in the Russian federation: clinical and diagnostic features. Med Parazitol (Mosk). 2013(1), 3–7 (2013)
4. Blanchette, M., Kunisawa, T., Sankoff, D.: Gene Order Breakpoint Evidence in Animal Mitochondrial Phylogeny. J. Mol. Evol. 49(2), 193–203 (1999)
5. Eds. Chauve, C., El-Mabrouk, N., Tannier, E.: Models and Algorithms for Genome Evolution. Comput. Biol. Series. London: Springer (2013)
6. Chen, Y.-L., Chen, C.-M., Pai, T.-W., Leong, H.-W., Chong, K.-F.: Homologous synteny block detection based on suffix tree algorithms. Journal of Bioinformatics and Computational Biology. 11(6), (2013)
7. Shao, M., Lin, Y., Moret, B.: Sorting genomes with rearrangements and segmental duplications through trajectory graphs. BMC Bioinformatics. 14(Suppl 15):S9 (2013)
8. Shao, M., Lin, Y., Moret, B.: An Exact Algorithm to Compute the Double-Cut-and-Join Distance for Genomes with Duplicate Genes. Journal of Computational Biology. (2014)
9. Wang, L.-S. and Warnow, T.: Distance-based genome rearrangement phylogeny. In: Mathematics of Evolution and Phylogeny, O. Gascuel, ed., Oxford Univ. Press, 353–380 (2005)
10. Wang, L.-S., Warnow, T., Moret, B.M.E., Jansen, R.K., and Raubeson, L.A.: Distance-based Genome Rearrangement Phylogeny. Journal of Molecular Evolution. 63(4), 473–483 (2006)
11. Moret, B.M.E., Wang, L.-S., Warnow, T., Wyman, S.: New approaches for reconstructing phylogenies based on gene order. Bioinformatics. 17(Supp), 165–173 (2001)
12. Baudet, C., Dias, U., Dias, Z.: Length and Symmetry on the Sorting by Weighted Inversions Problem. Advances in Bioinformatics and Computational Biology. Lecture Notes in Computer Science, 8826, 99–106 (2014)
13. Zverkov, O.A., Seliverstov, A.V., Lyubetsky, V.A.: Plastid-encoded protein families specific for narrow taxonomic groups of algae and protozoa. Molecular Biology. 46(5), 717–726 (2012)
14. Lyubetsky, V.A., Seliverstov, A.V., Zverkov, O.A.: Elaboration of the Homologous Plastid-Encoded Protein Families that Separate Paralogs in Magnoliophytes. Mathematical Biology and Bioinformatics. 8(1), 225–233 (2013) (in Russian)
15. Gorbunov, K.Yu., Gershgorin, R.A., and Lyubetsky, V.A.: Rearrangement and Inference of Chromosome Structures. Molecular Biology. 49(3), 327–338 (2015)
16. Zverkov, O.A., Seliverstov, A.V., Lyubetsky, V.A.: A database of plastid protein families from red algae and Apicomplexa and expression regulation of the moeB gene. BioMed Research International. 2014:510598 (2014)
17. Wei, L., Xin, Y., Wang, D., Jing, X., Zhou, Q., Su, X., Jia, J., Ning, K., Chen, F., Hu, Q., Xu, J.: Nannochloropsis plastid and mitochondrial phylogenomes reveal organelle diversification mechanism and intragenus phylotyping strategy in microalgae. BMC Genomics. 14:534 (2013)
18. Imanian, B., Pombert, J.F., Keeling, P.J.: The complete plastid genomes of the two 'dinotoms' Durinskia baltica and Kryptoperidinium foliaceum. PLoS ONE. 5(5):E10711 (2010)

19. Ong, H.C., Wilhelm, S.W., Gobler, C.J., Bullerjahn, G., Jacobs, M.A., McKay, J., Sims, E.H., Gillett, W.G., Zhou, Y., Haugen, E., Rocap, G., Cattolico, R.A.: Analyses of the complete chloroplast genome sequences of two members of the Pelagophyceae: Aureococcus anophagefferens CCMP1984 and Aureoumbra lagunensis CCMP1507. J. Phycol. 46(3), 602–615 (2010)
20. Cattolico, R.A., Jacobs, M.A., Zhou, Y., Chang, J., Duplessis, M., Lybrand, T., McKay, J., Ong, H.C., Sims, E., Rocap, G.: Chloroplast genome sequencing analysis of Heterosigma akashiwo CCMP452 (West Atlantic) and NIES293 (West Pacific) strains. BMC Genomics. 9:211 (2009)
21. Wang, X., Shao, Z., Fu, W., Yao, J., Hu, Q., Duan, D.: Chloroplast genome of one brown seaweed, Saccharina japonica (Laminariales, Phaeophyta): its structural features and phylogenetic analyses with other photosynthetic plastids. Mar. Genomics. 10, 1–9 (2013)
22. Le Corguille, G., Pearson, G., Valente, M., Viegas, C., Gschloessl, B., Corre, E., Bailly, X., Peters, A.F., Jubin, C., Vacherie, B., Cock, J.M., Leblanc, C.: Plastid genomes of two brown algae, Ectocarpus siliculosus and Fucus vesiculosus: further insights on the evolution of red-algal derived plastids. BMC Evol. Biol. 9:253 (2009)
23. Janouškovec, J., Horak, A., Obornik, M., Lukes, J., Keeling, P.J.: A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc. Natl. Acad. Sci. U.S.A. 107(24), 10949–10954 (2010)
24. Sadovskaya, T.A., Seliverstov, A.V.: Analysis of the 5'-Leader Regions of Several Plastid Genes in Protozoa of the Phylum Apicomplexa and Red Algae. Molecular Biology. 43(4), 552–556 (2009)
25. Baurain, D., Brinkmann, H., Petersen, J., Rodriguez-Ezpeleta, N., Stechmann, A., Demoulin, V. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Mol. Biol. Evol., 27(7), 1698–1709 (2010)
26. Lemieux, C., Otis, C., Turmel, M.: 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(2), 1–17 (2007)
27. Seliverstov, A.V., Lysenko, E.A., Lyubetsky, V.A.: Rapid evolution of promoters for the plastome gene ndhF in flowering plants, Russian Journal of Plant Physiology. 56(6), 838–845 (2009)
28. Lyubetsky, V.A., Zverkov, O.A., Rubanov, L.I., Seliverstov, A.V.: Modeling RNA polymerase competition: the effect of σ-subunit knockout and heat shock on gene transcription level. Biology Direct. 6(3) (2011)
29. Seliverstov, A.V., Lyubetsky, V.A.: Translation Regulation of Intron-containing Genes in Chloroplasts. Journal of Bioinformatics and Computational Biology. 4(4), 783–792 (2006)