J Mol Evol (2012) 75:119–129
DOI 10.1007/s00239-012-9523-2

Euglena gracilis and Trypanosomatids Possess Common Patterns in Predicted Mitochondrial Targeting Presequences

Katarína Krnáčová · Matej Vesteg · Vladimír Hampl · Čestmír Vlček · Anton Horváth

Received: 20 February 2012 / Accepted: 24 September 2012 / Published online: 12 October 2012

Abstract *Euglena gracilis* possessing chloroplasts of secondary green algal origin and parasitic trypanosomatids *Trypanosoma brucei*, *Trypanosoma cruzi* and *Leishmania major* belong to the protist phylum Euglenozoa. Euglenozoa might be among the earliest eukaryotic branches bearing ancestral traits reminiscent of the last eukaryotic common ancestor (LECA) or missing features present in other eukaryotes. LECA most likely possessed mitochondria of endosymbiotic α-proteobacterial origin. In this study, we searched for the presence of homologs of mitochondria-targeted proteins from other organisms in the currently available EST dataset of *E. gracilis*. The common motifs in predicted N-terminal presequences and corresponding homologs from *T. brucei*, *T. cruzi* and *L. major* (if found) were analyzed. Other trypanosomatid mitochondrial protein precursor (e.g., those involved in RNA editing) were also included in the analysis. Mitochondrial presequences of *E. gracilis* and these trypanosomatids seem to be highly variable in sequence length (5–118 aa), but apparently share statistically significant similarities. In most cases, the common (M/L)RR motif is present at the N-terminus and it is probably responsible for recognition via import apparatus of mitochondrial outer membrane. Interestingly, this motif is present inside the predicted presequence region in some cases. In most presequences, this motif is followed by a hydrophobic region rich in alanine, leucine, and valine. In conclusion, either RR motif or arginine-rich region within hydrophobic aa-s present at the N-terminus of a preprotein can be sufficient signals for mitochondrial import irrespective of presequence length in Euglenozoa.

Keywords Cleaved targeting sequence · Euglenids · Euglenozoa · Excavata · Kinetoplastids · Mitochondrial protein import · Protein motifs

Electronic supplementary material The online version of this article (doi:10.1007/s00239-012-9523-2) contains supplementary material, which is available to authorized users.

K. Krnáčová · A. Horváth
Department of Biochemistry, Faculty of Natural Sciences, Comenius University, 842 15 Bratislava, Slovakia

M. Vesteg
Department of Biology and Ecology, Faculty of Science, University of Ostrava, 710 00 Ostrava, Czech Republic

M. Vesteg
Department of Genetics, Faculty of Natural Sciences, Comenius University, 842 15 Bratislava, Slovakia

M. Vesteg (✉)
Department of Biology and Ecology, Faculty of Science, University of Ostrava, Chittussiho 10, 710 00 Ostrava, Czech Republic
e-mail: vesteg@fns.uniba.sk; matej.vesteg@osu.cz

V. Hampl
Department of Parasitology, Faculty of Science, Charles University in Prague, 128 43 Prague, Czech Republic

Č. Vlček
Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, 142 20 Prague, Czech Republic

Introduction

Euglenozoa is the protist phylum comprising euglenids, kinetoplastids (bodonids and trypanosomatids), and diplonemids. The support for grouping of these organisms has arisen from various morphological, molecular as well as phylogenetic evidences. These organisms possess discoidal mitochondrial cristae and characteristic feeding apparatus

(Simpson 1997; Triemer and Farmer 1991), they possess unusual base "J" in nuclear DNA (Dooijes et al. 2000), and they add non-coding capped spliced-leader (SL) RNA to nearly all cytosolic mRNAs via *trans*-splicing (Bonen 1993; Liang et al. 2003). The monophyly of Euglenozoa has been supported by molecular phylogenies as well (Simpson and Roger 2004).

The phylogenies support the early divergence of euglenids within Euglenozoa followed by the split of diplonemids and kinetoplastids (Simpson and Roger 2004; Simpson et al. 2002, 2004). However, the discovery of new euglenozoan clade—Symbiontida, exact phylogenetic position of which within Euglenozoa is currently uncertain (Breglia et al. 2010; Yubuki et al. 2009), has challenged this view, and diplonemids and symbiontids might be instead more closely related to euglenids than to kinetoplastids (Chan et al. 2012). Nevertheless, trypanosomatid parasites seem to be one of the latest branches of Euglenozoa, and they evolved most likely from one of the free-living bodonid clades (Callahan et al. 2002; Deschamps et al. 2011; Dyková et al. 2003; von der Heyden et al. 2004; Moreira et al. 2004; Simpson and Roger 2004; Simpson et al. 2004, 2006). The parasitism evolved at least four times within kinetoplastids (Simpson et al. 2006). Most euglenids are free-living heterotrophic flagellates, while some euglenids possess plastids of secondary green algal origin. Various lines of evidence suggest relatively recent acquisition of plastids by a phagotrophic ancestor of plastid-bearing euglenids (Leander et al. 2001; Leander 2004; Nozaki et al. 2003; Rogers et al. 2007; Turmel et al. 2009; Vesteg et al. 2010). Although the vast majority of phototrophic euglenids have already lost the ability of phagocytosis, marine phototrophic euglenid *Rapaza viridis* has been recently described, which has retained the ability to capture a eukaryotic prey (Yamaguchi et al. 2012). However, the flagellate *Euglena gracilis*, belonging to the latest euglenid clade represented by freshwater phototrophs (Linton et al. 2010), has been the most studied euglenid species so far. The scheme of euglenozoan phylogeny is presented in Supplemental Fig. 1.

Euglenozoan mitochondrial genome structures are unusual and diverse. The *E. gracilis* mitochondrial genome is represented by a heterodisperse collection of short molecules (approximately 4 kb) encoding gene fragments flanked by repeats (Spencer and Gray 2011). The diplonemid *Diplonema papillatum* possesses multiple 6–7 kb circular-mapping chromosomes containing short subgenic modules expressed as separate transcripts that are then *trans*-spliced to yield translatable mRNAs (Marande and Burger 2007; Vlček et al. 2011). In kinetoplastids, mitochondrial transcripts encoded by maxicircles are edited by guide RNAs (gRNAs) encoded by minicircles (Hajduk et al. 1993), although in some trypanosomatids a small proportion of gRNAs is also encoded by maxicircles (for review see Simpson et al. 2000; Stuart and Panigrahi 2002). The RNA editing in mitochondria of kinetoplastids includes uridine insertions and deletions.

Euglenozoa have been classified within Excavata which have been only recently suggested to be one of the three eukaryotic major groups possibly representing the most basal eukaryotic branch (Hampl et al. 2009). However, another hypothesis has been recently suggested proposing that Euglenozoa might be instead the earliest branching eukaryotes apart from Excavata (Cavalier-Smith 2010). Cavalier-Smith (2010) has proposed that some euglenozoan features are primitive. Two of these include euglenozoan mitochondrial features—unique cytochrome *c* biogenesis (Allen et al. 2008) and the possible absence of mitochondrial outer membrane channel Tom40 (Schneider et al. 2008).

The engulfment of an α-proteobacterial ancestor of mitochondria by a host entity was probably a key moment in eukaryogenesis (Martin and Müller 1998; Vesteg and Krajčovič 2008, 2011). The α-proteobacterial ancestors of mitochondria might have been either strict aerobes (Cavalier-Smith 2002) or facultative anaerobes (Martin and Müller 1998). The former view is supported by the fact that most eukaryotes possess aerobic mitochondria and by the calculation of an oxyphobic index considering the amino acid distribution in anaerobes and aerobes suggesting that the last eukaryotic common ancestor (LECA) was an aerobe (Di Giulio 2007). The latter hypothesis is supported by the fact that mitochondria of *E. gracilis*, other euglenozoans and excavates possess biochemical properties of both aerobic and anaerobic mitochondria (Ginger et al. 2010), and thus, the biochemistry of mitochondria of these organisms might represent an intermediary evolutionary stage reminiscent of the mitochondria of LECA.

The acquisition of an α-proteobacterium and its evolution to a primitive mitochondrion was accompanied by the transfer of endosymbiont genes to the host genome and the evolution of a mechanism for import of proteins to mitochondria including the evolution of mitochondria-targeting presequences. The most of proteins necessary for mitochondrial function were probably nucleus-encoded in LECA (Desmond et al. 2011). The potentially primitive mitochondrial import apparatus in possibly most ancient eukaryotic group (either euglenozoans or excavates) could be a good model to trace the evolution of mitochondrial import mechanism of the first eukaryote. However, the data about mitochondrial targeting presequences of Euglenozoa are fragmentary.

The studies of some proteins targeted to mitochondria and some predictions suggest that mitochondrial presequences of trypanosomatids are quite short (some only 6 aa in length) (Häusler et al. 1997). Nevertheless, e.g., the trCOIV (Cox4) preprotein of *Leishmania tarentolae*

possesses 31 aa-long presequence (Maslov et al. 2002). In contrast to other eukaryotes, cytochrome c_1 lacks cleaved targeting peptide in euglenozoans (Priest et al. 1993; Priest and Hajduk 2003). Experimental evidence exists that *E. gracilis* presequences of the subunits II and IX of ubiquinol-cytochrome *c* reductase complex (Qcr2 and Qcr9) are 42 and 30 aa-long, respectively (Cui et al. 1994). Although 30 aa-long consensus sequence has been generated from the N-termini of 107 hypothetical *E. gracilis* proteins potentially targeted to mitochondria (Gawryluk and Gray 2009), nearly nothing is known about the variability of length of mitochondrial presequences in euglenids. While the structure of plastid-targeting presequences and domains and motifs therein in *E. gracilis* have been precisely analyzed (Durnford and Gray 2006), the common patterns present in mitochondrial presequences of euglenids are largely unknown.

In this study, we searched for common protein motifs in predicted mitochondrial presequences of nucleus-encoded mitochondrial precursor proteins in *E. gracilis* and parasitic trypanosomatids *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major*. Since trypanosomatid parasites and phototrophic freshwater euglenids (including *E. gracilis*) seem to be the most distant branches in euglenozoan phylogeny (see Supplemental Fig. 1), the similarities of mitochondrial import signals and machineries in these organisms could likely reflect the nature of mitochondrial import apparatus of euglenozoan common ancestor, if not LECA itself.

Methods

More than 500 proteins of respiratory chain and associated proteins, proteins of citric acid cycle, and proteins involved in the synthesis of Fe–S clusters from *T. brucei*, *Chlamydomonas reinhardtii*, *Saccharomyces cerevisiae* and *Bos taurus* were used as queries in tBLASTn search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) in currently available non-annotated EST data (http://www.ncbi.nlm.nih.gov/) of *E. gracilis* (Ahmadinejad et al. 2007; Durnford and Gray 2006; Ferreira et al. 2007). In addition, the *E. gracilis* EST data were searched for the presence of homologs encoding 20 trypanosomatid editosomal proteins. Only those hits with e-values below 1×10^{-9} were used for further analysis. Some of these contained SL-leader sequence TTTTTTTCG generally present at the 5'-end of *E. gracilis* mRNAs and mRNAs of other euglenids (Frantz et al. 2000). The protein sequences used for analysis of presequence-encoding regions were either those obtained via in-frame translation (based on comparison with homologous proteins from other organisms) of ESTs containing SL-leader or sequences obtained via linking of 2–5 EST sequences (at least one containing SL-leader) with identical overlaps. The first methionine encoded by ATG in the sequence downstream of SL-leader was chosen as the presequence start. Accession numbers of *E. gracilis* ESTs used in this study, the names of the organisms with the best BLASTx hits, e-values and names of putative gene products are in the Supplemental Table 1. Mitochondrial protein homologs of *T. brucei* (TB), *T. cruzi* (TC), and *L. major* (LM) (if found) were also included in the analysis of presequence regions, as were the mitochondrial proteins involved in heme synthesis ferrochelatase (FeCH), δ-aminolevulinic acid synthase (ALAS) and protoporphyrinogen oxidase (PPOX) in *E. gracilis* (Kořený and Oborník 2011). In addition, various other *T. brucei* (TB), *T. cruzi* (TC), and *L. major* (LM) mitochondrial protein precursors, corresponding homologs of which were not found in currently available *E. gracilis* sequence data, involved in citric cycle, synthesis of Fe–S clusters and RNA editing were also used in this study. Accession numbers (http://www.ncbi.nlm.nih.gov/) and names of trypanosomatid nucleus-encoded mitochondrial protein precursors included in the analysis of euglenozoan mitochondrial presequences can be found in the Supplemental Table 2.

E. gracilis protein sequences and trypanosomatid proteins listed in Supplemental Tables 1 and 2 were used as queries in tBLASTn search in currently available ESTs of *Euglena longa*, Bodonidae, and Diplonemida, and our unpublished EST data of *Eutreptiella gymnastica*. The currently available EST data of *E. gracilis* and *E. gymnastica* were also screened for the presence of Tom40—key component of the complex of mitochondrial outer membrane translocon. The screening was performed, firstly, using tBLASTn with Tom40 (*S. cerevisiae*) and ATOM (*T. brucei*) as queries and, secondly, using Hidden Markov model (HMM) search. The HMM search (Likic et al. 2010) was performed using MyHMMER script kindly provided by Vojtěch Žárský (Department of Parasitology, Faculty of Science, Charles University in Prague) under the default setting using alignments of either 23 eukaryotic Tom40 sequences or ATOM of kinetoplastids. The top hits from the searches were further evaluated using HHpred (Söding 2005) (http://hhpred.tuebingen.mpg.de/hhpred) under default setting and with pdb70_9Feb12 database. This database consisted of HMMs created from alignment of proteins present in protein data bank on February 9, 2012 (Bourne et al. 2004).

The probable cleaved targeting sequences (CTS) were identified using the programs MITOPROT (Claros and Vincens 1996) (http://ihg.gsf.de/ihg/mitoprot.html) and targetP (Emanuelsson et al. 2000; Nielsen et al. 1997) (http://www.cbs.dtu.dk/services/TargetP/) under default settings. When *E. gracilis* sequences were used for CTS prediction by targetP, organism group was changed from

non-plant to plant to consider potential plastid localization. Common sequence motifs within presequences were determined using MEME program (Bailey and Elkan 1994) (http://meme.nbcr.net/meme4_6_1/cgi-bin/meme.cgi) and GLAM2 (Frith et al. 2008) (http://meme.nbcr.net/meme4_6_1/cgi-bin/glam2.cgi). The biochemical properties of predicted CTSs such as molecular weights, theoretical isoelectric points (pI), and numbers of positively and negatively charged amino acids were calculated using +ProtParam program (Gasteiger et al. 2005) (http://expasy.org/tools/protparam.html).

Results

The predicted presequences (cleaved targeting sequences—CTSs) of mitochondrial protein precursors of *E. gracilis*, *T. brucei*, *T. cruzi* and *L. major* were analyzed in this study. It was not possible to find CTS in all proteins included in this study, although they were predicted to be targeted to mitochondria. For example, no CTS regions were identified in cytochromes c_1 and c of *E. gracilis* and all three trypanosomatids. Similarly, it was impossible to detect CTS in *E. gracilis* NADH dehydrogenase NDUFA9 and aconitase (Aco). The total number of euglenozoan proteins predicted to be targeted to mitochondria and possessing CTS identifiable by either MITOPROT or targetP programs (or both) was 127. While only 105 CTSs were detected by MITOPROT, 122 CTSs were detected by targetP. Therefore, CTSs predicted by targetP were chosen for further analysis.

The euglenozoan-predicted CTSs were 5–118 aa-long with an average size of CTS 31 and 41 aa predicted by MITOPROT and targetP, respectively (see Table 1 for details). The CTSs contained mainly positively charged and hydrophobic aa-s, and the average isoelectric point was 11.37 ± 1.26 (MITOPROT) and 10.66 ± 1.84 (targetP). The CTS-lengths and biochemical properties of presequences predicted by MITOPROT and targetP are depicted in Supplementary Table 3. In addition, *E. gracilis* Qcr1 respiratory chain protein was found to be homologous to β subunits of matrix processing peptidases (MPPs) from trypanosomatids. This was consistent with the previous study of Cui et al. (1994). Qcr1 and its partial homolog Qcr1p were found to be homologous to β subunits of different trypanosomatid MPPs (assigned B1 and B2 here). Supplemental Table 1 contains the list of *E. gracilis* mitochondrial preproteins included in this study, ESTs of which were found in the currently available non-annotated EST sequence data of this organism.

Since CTSs needed to be longer than 7 aa to be analyzed using MEME program such as to search for common motifs, 116 presequences predicted by targetP (each longer than 7 aa) were analyzed using this program. This analysis generated short conserved motif (M/L)RR present in most of the analyzed presequences. The consensus logo generated from all 116 proteins included in the analysis is presented in Fig. 1a. This motif is generally followed by hydrophobic region rich in alanine, phenylalanine, leucine, and valine. The short (M/L)RR motif is the same as previously identified at the N-terminus of some trypanosomatid mitochondrial presequences predicted to be short (Häusler et al. 1997). In the MEME output including all 116 presequence, the consensus region serving as the basis for generation of the logo was present at the N-terminus (up to 7 aa from the first methionine) in 60 euglenozoan presequences, while it was found somewhere within the predicted presequence region of 56 euglenozoan presequences (more than 7 aa from the N-terminus) (data not shown). The MEME analysis of 60 proteins in which the (M/L)RR motif was detected at the N-terminus in the initial step is presented in Table 2, and the common 10 aa-long motif (present in 56 of 60 of these proteins) is depicted in Fig. 1b. The MEME analysis of 56 proteins, in which the (M/L)RR motif was detected within the presequence region in the initial step, is presented in Table 3 and the common 20 aa-long logo generated from 50 of 56 of these proteins is depicted in Fig. 1e. Table 4 includes only the analysis of mitochondrial protein precursors from *E. gracilis* and MppB2 from *Eutreptiella gymnastica*, and 11 aa-long logo generated by the program is depicted in Fig. 1c. This logo is very similar to the N-terminus of the logo generated by

Table 1 Average, modus, and median length of euglenozoan presequences predicted by MITOPROT and targetP

Organism	MITOPROT				targetP			
	Range	Average	Modus	Median	Range	Average	Modus	Median
E. gracilis	12–64 aa	32 aa	22 aa	26 aa	5–96 aa	31 aa	23 aa	22 aa
T. brucei	11–54 aa	27 aa	11 aa	24.5 aa	6–116 aa	44 aa	15 aa	30 aa
T. cruzi	10–90 aa	28 aa	13 aa	19 aa	5–110 aa	44 aa	18 aa	34 aa
L. major	11–106 aa	36 aa	16 aa	28 aa	9–118 aa	44 aa	16 aa	29 aa
T. b. + T. c + L. m	10–106 aa	30 aa	13 aa	24 aa	5–118 aa	44 aa	16 aa	31.5 aa

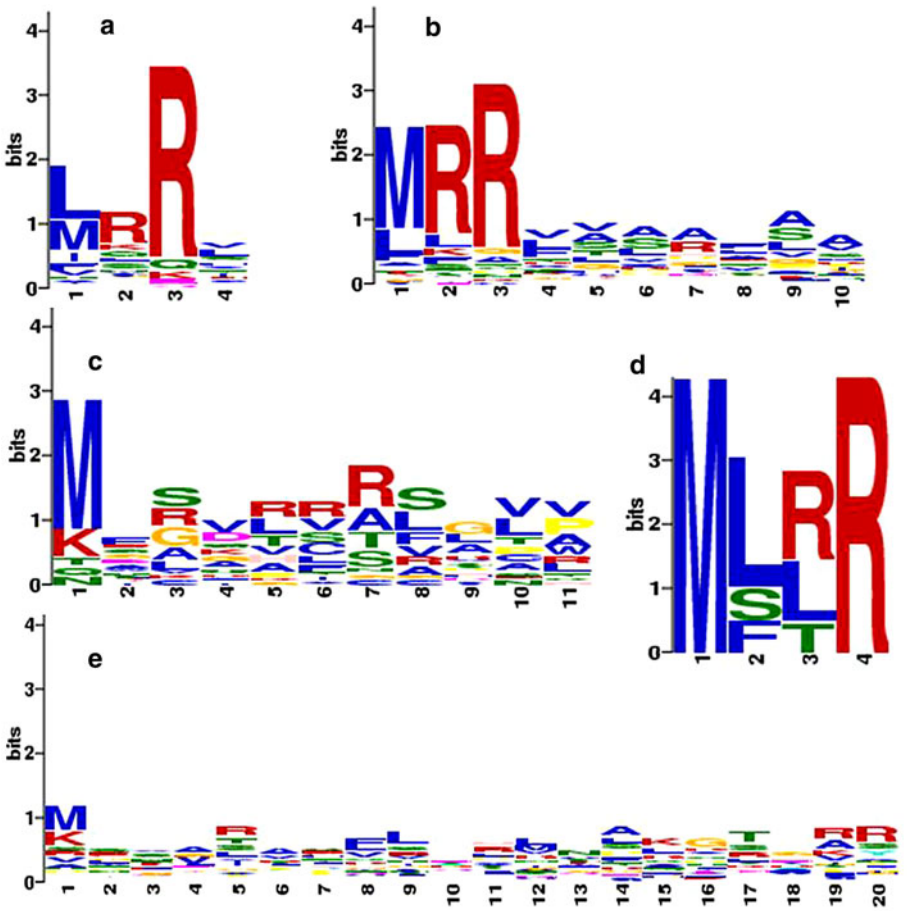

Fig. 1 Common protein motifs present in presequences of mitochondrial precursor proteins of *Euglena gracilis* and trypanosomatids. **a** Logo generated in the MEME analysis of 116 euglenozoan presequences predicted by targetP. **b** 10 aa-long logo generated in the MEME analysis of 60 euglenozoan mitochondrial presequences predicted by targetP in which the (M/L)RR motif was detected at the N-terminus (up to 7 aa from the first methionine) (see Table 2). **c** The 11 aa-long logo generated in the MEME analysis of 20 *E. gracilis* targetP-predicted mitochondrial presequences and 1 presequence from *Eutreptiella gymnastica* (see Table 4). **d** Logo generated in GLAM2 program from 6 euglenozoan sequences predicted to be 5 or 6 aa-long by TargetP. **e** 20 aa-long logo generated in the MEME analysis of 56 euglenozoan mitochondrial presequences (predicted by targetP) in which the (M/L)RR motif was detected more than 7 aa from the first methionine (see Table 3)

Gawryluk and Gray (2009). Figure 1d includes the common motif generated from six euglenozoan sequences predicted to be 5 or 6 aa-long by targetP. The sequence identities among presequences and among mature parts of preproteins in *E. gracilis* and trypanosomatids are shown in Supplemental Fig. 2.

In addition, we could not identify the mRNAs of proteins analyzed in this study in the currently available EST data of diplonemids. We could identify some of these in bodonid data, but none of these was predicted to be targeted to mitochondria by the programs. We identified ESTs encoding Cit1, Fh, Mdh, Grx, Hel61, MEAT1, trCOIV, AtpB, AtpA, CytC1, CytC, SdhB1, TAO, SdhB2, NuoG, MppB1 (Qcr1), MppB2 (Qcr1p), Eao2, and Eao3 in *Euglena longa* EST data. However, most of these lack 5'-ends. ESTs encoding cytochromes *c* and c_1 do possess 5'-end, but similarly, in *E. gracilis* and trypanosomatids, the CTSs were not detectable by the programs, because these proteins probably lack it. The only sequences encoding homologs of proteins analyzed in this study found in *E. gymnastica* transcriptome data (unpublished) were trCOIV, AtpA, AtpB, Mdh, MppB2, cytochromes *c* and c_1, while only MppB2, and cytochromes *c* and c_1 possessed complete 5'-end. While cytochromes *c* and c_1 lacked identifiable CTS, MppB2 presequence was predicted to be 40 aa-long by targetP (and 39 by MITOPROT).

Using BLAST and HMM search, we have identified a single eukaryotic porin3 protein in *E. gracilis* (accession number AF317222 in GenBank nr database and ELE00007502 in GenBank EST database), and its ortholog was detected also in the transcriptome of *E. gymnastica*. HHpred search indicated that this porin likely belongs to the VDAC (Voltage-dependent anion channel) subfamily rather than to Tom40 subfamily. No porin similar to ATOM of kinetoplastids has been detected in euglenids.

Discussion

The nucleus-encoded mitochondrial precursor proteins in *E. gracilis* and trypanosomatids apparently possess presequences sharing common features. Principally, the euglenozoan mitochondrial presequences are of three types with respect to the length. Two types of presequences were known in trypanosomatids before: (1) the short (up to 10 aa-long) mitochondrial targeting presequences, and (2) 10–30 aa-long presequences (most frequently about 16 aa-long) (Häusler et al. 1997). In this study, it has been

Table 2 The output of the MEME analysis of 60 euglenozoan mitochondrial presequences (predicted by targetP) in which the (M/L)RR motif was detected at the N-terminus (up to 7 aa from the first methionine) in the initial analysis in which 116 predicted euglenozoan mitochondrial targeting presequences were included

Protein-organism	CTS start	*p*-value	Sites		CTS length
AtpA-TC	0	2.69e-08	MRRFFSKFAA	GLPARF	16
AtpA-TB	0	8.62e-08	MRRFGSKFAS	GLASRC	16
MP42-TC	0	1.07e-07	MRRIASHLSQ	RTYRALFPRG	42
AtpA-LM	0	2.91e-07	MRRFVAQYVA	PAMGRL	16
MP46-TC	0	2.72e-06	MRRVSAHGVS	VHFAVVLTTA	26
NuoG-EG	0	3.11e-06	MRRVLARFGP	HVPRSFHTTV	23
NuoG-TB	0	3.11e-06	MRRFSNCLLC	FGAITPVGDS	31
MP42-LM	0	4.56e-06	MRRIGVATCQ	RRWL	14
MP42-TB	0	6.52e-06	MKRVTSHISR	FLPLVLSQRG	30
HEL61-LM	0	1.13e-05	MRRFASALFG	RWGAAQCGVV	104
SCO1/SCO2-TB	0	1.85e-05	MRRVSGNVLC	GSASHWRAEL	55
MP63-TC	0	2.23e-05	MRRLLLAGAL	AKFSRRVGAR	53
MP44-TC	0	2.68e-05	MKRGVLAFCV	AAPAAARC	18
Cox15-LM	0	3.19e-05	MQRFTARYVV	SAAASASARR	37
NuoG-TC	0	3.19e-05	MKRLAARFPL	WYAAPPLLMA	30
Fxn-TB	0	3.48e-05	MRRTCCATTS	AVLRSLVYLR	46
REL1-LM	0	3.79e-05	MRRLALRRAP	RCSHATL	17
Idh-LM2	0	5.75e-05	MFRHVSAASA	SSLVAARSFS	26
Idh-TB1	1	9.21e-05	M FRRVACAASS	VNAGALTPRF	29
Cit1-TC	0	1.54e-04	MMRFCARAGI	LLNAPSAARR	36
MP44-LM	7	2.51e-04	MKRTCRS LRRFSAGPLD	LAHSICRSAV	34
Csd-LM	3	2.51e-04	MRC MSRVFLCAAA	STANGAGAAS	95
FeCH-EG	0	2.87e-04	MERFFHAAAS	WKGVGLALST	42
MP61-LM	6	2.87e-04	MNRCGV KRRLSLRLPV	V	17
Idh-TC2	1	3.07e-04	M FRRVACRVPF	TAAAVCY	18
MppB1-LM	0	3.28e-04	MLRRTSAVAA	TAALPHNMTM	14
MP67-LM	2	3.28e-04	MR LRRTSAWRAV	CHIPDALRTH	104
Csd-TC2	0	3.28e-04	MFRGVCGLLG	AAKATPTAAA	42
SdhB2-TB	1	3.74e-04	M LRKVTSPYKV	SIRR	15
SdhB2-LM	0	4.84e-04	MLRKVAPRPY	KTMVRR	16
MP81-TB	0	5.16e-04	MRRLTRRSGR	LS	12
MP44-TB	0	7.04e-04	MRRAVVLRTA	AP	12
MEAT1-LM	3	7.95e-04	MQA GRRKLVAESI	AAYRA	18
trCOIV-EG	0	8.45e-04	MLRQVVRRSN	PLRMQVRG	18
MP48-TB	1	9.52e-04	M LRRLGVRHFR	RTPLLF	17
MP18-LM	1	1.01e-03	M FRRLSPAAAT	RSM	14
REL2-TC	1	1.07e-03	M LRRHFQLFLR	RT	13
MP100-TC	2	1.07e-03	MA LSRTWCRFAV	TTYRQS	18
MP46-TB	10	1.13e-03	MLRVENLRRS MTRLARHSLI	RAVPFSPLSV	87
NDUFV1-TB	0	1.59e-03	MLRRVGFLSA	TGSLL	15
RET2-TC	44	1.99e-03	AWGKAILTEN YRRVGPEHMF	RTAIRA	60
MP99-TB	1	1.99e-03	M LRRSRLHLLA	DYRT	15
MP100-LM	0	2.59e-03	MRGALARSAC	RL	12
MP63-TB	22	3.35e-03	HLGTTSAQCL MRATKYPCGA	MCRN	36
MP18-TB	15	3.71e-03	SRRLLLQQTF MRCCKSVNSV	TLVGVVHDIQ	75
AtpB-LM	8	3.89e-03	MLSRVQSA MIRRAAGVRA		18
trCOIV-TB	2	3.89e-03	MF ARRSLIATVA	AA	14
MppB1-TB	0	4.97e-03	MFRPSFCRCL	PVLNCTLSAP	25
AtpB-TB	9	4.97e-03	MLTRFRSAV LRGAVSITGA	RA	21
Cox15-TB	23	6.29e-03	GWSRGPALWS TRRLESMGST	SWRVPSTNKA	109
REL2-LM	0	9.43e-03	MLRRCLLTRL	VRRSLVLF	18
Idh-EGB	0	1.22e-02	MRAVLRSAAL	A	11
NDUFA9-TB	6	1.27e-02	MSRRVF ARNMNGEISS	TWRGGGSEAN	72
AtpB-EG	7	1.76e-02	MQAIRRS LRTVAKPMVG	RMM	20
Eao2-EG	9	2.06e-02	MALSRVASL CRPMGAGPSP	LWLLRAY	26
trCOIV-TC	3	3.11e-02	MLS RRSLTTAFAA	M	14

CTS Cleaved targeting sequence; *EG E. gracilis*; TB *T. brucei*; TC *T. cruzi*; LM *L. major* (for the description of proteins, see Supplemental Tables 1 and 2, and "Methods" section). The common 10 aa-long motif (logo) generated from 56 of 60 proteins included in the analysis is depicted in the Fig. 1b

Table 3 The output of the MEME analysis of 56 euglenozoan mitochondrial presequences (predicted by targetP) in which the (M/L)RR motif was detected more than 7 aa from the first methionine in the initial analysis in which 116 predicted euglenozoan mitochondrial targeting presequences were included

Protein-organism	CTS Start	*p*-value	Sites			CTS length
MppB1-EG	63	9.50e-20	IDAGSRWETE	KNNGVAHFLEHMNFKGTGKR	S	84
MppB2-EG	63	9.50e-20	IDAGSRWETE	KNNGVAHFLEHMNFKGTGKR	SRQDIEFGME	94
MppB1-TC	59	1.11e-18	IDAGSRFEDL	RNNGVAHFLEHMNFKGTEKY	SKRA	93
REL1-TC	80	1.33e-07	LGAQDWVACE	KVHGTNFSIYLINLGDKEVV	RFAKRSGIMD	110
HEL61-TC	0	2.59e-07		MRVAIQMNGIHLKLLRSCRR	PIHNTFENAE	44
MEAT1-TB	0	2.59e-07		MRVAIQMNGIHLKLLRSCRR	PIHNTFENAE	44
MP52-TB	86	2.59e-07	LAAQEWVACE	KVHGTNFGIYLINQGDHEVV	RFAKRSGIMD	116
SdhB2-TC	1	4.93e-07	M	NICALLFLSYFCDGKGSHAR	VHIRAQTPTQ	70
SdhB2-EG	0	7.06e-07		MSVVRRCFQRALRPLGQRAY		20
Mdh-TC2	13	7.06e-07	RGITPTNWLL	KEKEMVFFLRRVAPKKTSGK	VVVFGATTVV	80
Aco-LM	44	5.00e-06	RNCDEFDVTS	KTVESIFDWKDNCTKGIEIP	FKPARVVLQD	89
Mdh-LM2	54	5.88e-06	CCNTAADDDV	PGSGIAADLSHIDTLPKVHY		74
HEL61-EGC	48	5.64e-05	KQVCASAATG	GGKTAAFVLPILHTLAQDPY	GVYAVIVTPS	80
EtfQ-LM	49	8.56e-05	LSAAIRLKQL	AGDQRDSFRVGLVEKGSEIG	AHTLSGACVE	118
Idh-EG	1	1.28e-04	M	KAVHSRLFCSPVSRLALQRW	Y	22
Idh-EGA	0	1.47e-04		MAGGLFRCLSVTCAQSSLRR	SMRH	24
EtfQ-EG	0	1.67e-04		MISRRVRLCLPRLGNILYRS	Y	21
SCO1/SCO2-TC	12	1.67e-04	RRGWGLLAAS	HTTRRVTITALWCAKRKYGR	MWEPTCGTRR	74
MP99-TC	15	2.31e-04	RILWTADFRT	SPKISEEFNNVLNRSGWGFD	HVAIDMGALA	99
MP61-TB	26	2.62e-04	LPSVSHTLRC	RGTISNSGGEGNDPIAEPRR	VRLVPSAPHG	106
SCO1/SCO2-LM	7	3.37e-04	MSEAVDE	MRDNPVWMLWALGFLTLGVV	TVVIS	32
MP24-TC	62	4.58e-04	PTFVADPMRP	RQQIGMDGSDYCNERARDRI	RFASRC	88
TAO-TB	59	4.58e-04	VPLRVSDESS	EDRPTWSLPDIENVAITHKK	PNGLVDTLAY	91
Cox15-TC	0	5.49e-04		MLRFRPRVFQTNTTRLHWAF	LFKRRLQSTV	33
MP18-TC	52	5.49e-04	AVTQFTLTTT	SIDTTHPTQEVVVEKDHHTI	RCF	75
MP90-TC	0	6.55e-04		MRLVGSCVRSRGLLWSEWQV	TWRLCRHF	28
Mdh-EGA	0	6.95e-04		MFSKTSSFVPLGRAFSTTRS	QN	22
MP46-LM	52	7.80e-04	ADVIAPRLSF	AGNRRATVFTSSCCRLTPAR	C	73
MP63-LM	14	8.76e-04	RRIGVLAAAA	ASRCALRRLHHVSSRPLHAS	GAVASAKGAS	70
HEL61-TB	10	8.76e-04	MRALRCVRRG	VYRQSVRLCYFMSLECSLRP	ITGASARLLC	78
PPOX-EG	26	9.27e-04	GLATAYYLRE	VPGVRVTVLEATAAPGGWCR	TLPQPGPSWM	96
MppB2-EUT	0	1.04e-03		MSAAPPLSSILRHSKPVFKQ	SLKTASPVLQ	40
Mdh-TB2	22	1.10e-03	VTGKVVVFGA	TTNVGKHLSLLLTLSPQVKE	LCCFDPLNDV	92
Aco-TC	9	1.23e-03	ADGGEAKYF	KLHEIDPRYETLPFSIRVLL	ESAVRNCDEF	88
MP90-LM	2	1.30e-03	MP	TSSLHLSYTQAMHSSATHRL		22
MP57-TB	0	1.45e-03		MLMHTAPWLHMRLSRLFRQS	PLSL	24
MP67-TC	33	2.73e-03	TDIFSWGSSP	PEQDGGVSKGEMNRMKSRFY		53
MppB2-TC	0	2.88e-03		MSVSFSTLVQRARPASNHAT	TQALRS	26
MP24-TB	3	3.02e-03	MRV	RSLLLCTRRDPLQRAVDVAY	ASGMLLGSGS	53
MP24-LM	23	3.52e-03	AGCFDSANGS	SSSLLRRCTDSACAQGRNVT	S	44
SCO1/SCO2-EG	1	4.49e-03	M	WRGALRSALRPPGAWRRAAR	PS	23
Fxn-LM	37	6.25e-03	SHSFANASTS	VMTASAMAVVRRAATTTGAS	A	58
MP90-TB	0	6.55e-03		MGLHWPLARSSNWCHMRALS	NEHLRRRVGA	42
Cox15-EG	0	7.51e-03		MPLCVAVGVMRHGLRGALPR	TAALLPCRPL	49
Idh-EGC	0	1.02e-02		MRAARCLLRTFRIAVNTGDG	IGADVVPAAV	44
trCOIV-LM	0	1.12e-02		MLTRRAVSSAVGAAMVTSSS	VSMQRRYDHD	33
RET2-LM	64	1.56e-02	NDHYVQWGRA	LLEENSKRIGPEQMFRTAIR	A	86
Cit1-TB	0	1.75e-02		MCMRARYSSGIVRGAMLRGF	SSSPSLF	27
Mdh-EGD	3	2.30e-02	MFL	KASASLAPLGRAFSTTRGCN		23
SdhB1-TC	2	3.94e-02	MP	SAPLTGEVARYSSPLFMYRR		22

CTS Cleaved targeting sequence; EG *E. gracilis* (EGA—EGD: different homologs of the enzymes); TB*T. brucei*; TC *T. cruzi*; LM *L. major*; EUT *Eutreptiella gymnastica* (for the description of proteins, see Supplemental Tables 1 and 2, and "Methods" section). The common 20 aa-long motif (logo) generated from 50 of 56 proteins included in this analysis is depicted in the Fig. 1e

demonstrated that some *E. gracilis* mitochondrial precursor proteins also possess these two types of CTS. Moreover, this study revealed that some mitochondial targeting presequences can be quite long (up to 118 aa) in both *E. gracilis* and trypanosomatids. These predictions are consistent with the experimental evidence that frataxin precursor of *T. brucei* possesses 55 aa-long mitochondrial targeting sequence (Long et al. 2008), and that 115 aa-long N-terminus of WD-repeat preprotein can serve as mitochondrial targeting signal in *T. cruzi* (Bromley et al. 2004). Most of long euglenozoan presequences identified here possess at least one long common protein motif (up to 20 aa) (Fig. 1e). The variability of mitochondrial presequence length in *E. gracilis* was also predicted via the alignments of eukarytic enzymes involved in tetrapyrrole synthesis and their bacterial homologs lacking CTS (Kořený and Oborník 2011). The presequences of mitochondria-targeted enzymes ferrochelatase (FeCH) and

Table 4 The output of the MEME analysis of 20 *E. gracilis* targetP-predicted mitochondrial presequences

Protein-organism	CTS start	*p*-value	Sites			CTS length
MppB2-EG	42	1.63e-11	PNGFRIASES	KDGDTCTVGVW	IDAGSRWETE	94
MppB1-EG	42	1.63e-11	PNGFRIASES	KDGDTCTVGVW	IDAGSRWETE	84
Mdh-EGA	0	1.65e-05		MFSKTSSFVPL	GRAFSTTRSQ	22
Idh-EG	0	1.81e-05		MKAVHSRLFCS	PVSRLALQRW	22
Mdh-EGD	0	4.15e-05		MFLKASASLAP	LGRAFSTTRG	23
EtfQ-EG	0	7.58e-05		MISRRVRLCLP	RLGNILYRSY	21
AtpB-EG	0	8.23e-05		MQAIRRSLRTV	AKPMVGRMM	20
Idh-EGC	15	9.70e-05	CLLRTFRIAV	NTGDGIGADVV	PAAVRVLELV	44
Idh-EGA	0	1.33e-04		MAGGLFRCLSV	TCAQSSLRRS	24
NuoG-EG	0	1.56e-04		MRRVLARFGPH	VPRSFHTTVS	23
HEL61-EGC	36	1.82e-04	VQKAAIPLIL	QGKQVCASAAT	GGGKTAAFVL	80
Eao3-EG	0	2.45e-04		MVCAHLTRHVA	LRGLPRTM	19
AtpA-EG	0	2.63e-04		MFSTLRNGRLV	ARN	14
SCO1/SCO2-EG	0	3.26e-04		MWRGALRSALR	PPGAWRRAAR	23
MppB2-EUT	18	3.75e-04	SILRHSKPVF	KQSLKTASPVL	QSSLGNGFRV	40
trCOIV-EG	0	3.75e-04		MLRQVVRRSNP	LRMQVRG	18
Idh-EGB	0	4.93e-04		MRAVLRSAALA		11
Eao2-EG	0	5.27e-04		MALSRVASLCR	PMGAGPSPLW	26
SdhB2-EG	0	8.28e-04		MSVVRRCFQRA	LRPLGQRAY	20
Cox15-EG	33	2.36e-03	LLPCRPLGRA	TGRAPFRLGTP	IEMRH	49

CTS Cleaved targeting sequence; EG *E. gracilis* (EGA—EGD: different homologs of the enzymes); EUT *Eutreptiella gymnastica* (for the description of proteins, see Supplemental Tables 1 and 2, and "Methods" section). The common 11 aa-long motif (logo) generated by this analysis is depicted in the Fig. 1c

δ-aminolevulinic acid synthase (ALAS) seem to be quite long (47 and 86 aa, respectively), while the presequence of protoporphyrinogen oxidase (PPOX) may be only 7 aa-long (see Supplemental Fig. S3 in Kořený and Oborník 2011). While ALAS mitochondrial presequence was not detectable by both MITOPROT and targetP, FeCH, and PPOX presequences were predicted to be 42 and 96 aa-long by targetP (and 49 and 26 aa-long by MITOPROT) in our analysis, respectively. This suggests that the usage of the programs for the prediction of mitochondrial presequences is not without its limitation.

Although the length of a presequence predicted by MITOPROT and targetP differed in many cases (Supplemental Table 3), most of predicted presequences contained arginine (R) as the last but one aa (data not shown). This may reflect the fact that these programs search for conserved cleavage sites most commonly found among eukaryotes, while their algorithms differ. It is also possible that euglenozoan processing peptidases do not recognize consensus cleavage sites in some cases. It should be mentioned that both MITOPROT and targetP are mainly trained with model organisms which are phylogenetically distant from euglenozoans. Although the cleavage sites of most euglenozoan presequences included in this study remain to be experimentally verified, the MEME analysis revealing common motifs at the N-termini of predicted presequences is of significance. Since the analyzed mitochondrial proteins have different mitochondrial function such as in oxidative phosphorylation, citrate cycle, heme synthesis, synthesis of Fe–S clusters, RNA editing, and RNA and protein processing (Supplemental Tables 1 and 2), the common motif identified at the N-termini of most euglenozoan presequences could be hardly explained by the redundancy of the data used in this study.

Another possibility for the explanation of the difference between presequence lengths predicted by targetP and MITOPROT is that some precursor proteins are sequentially processed to the mature form in two steps by peptidases, while each program suggests only one of the cleavage sites. Two-step processing of protein precursor has been demonstrated experimentally, e.g., for Rieske iron-sulfur protein subunit of the cytochrome *c* reductase in *T. brucei* (Priest and Hajduk 1996). Moreover, even more complex processing of frataxin in *T. brucei* occurs (Long et al. 2008). Since we found two different homologs of *E. gracilis* Qcr1 (possessing domains typical for β subunits of MPPs) as well as two homologs of MPP β subunits in trypanosomatids, it might be possible that two types of MPPs exist in Euglenozoa probably recognizing different cleavage sites. Hypothetically, cleavage sites might alternate under different growth conditions or in different developmental stages in the case of trypanosomatids. It has been shown for example that alternative oxidase (TAO) and subunit IV of cytochrome oxidase (trCOIV) are imported in the mitochondrion via different mechanisms in *T. brucei* bloodstream and procyclic form (Williams et al. 2008).

Euglenozoan mitochondrial import apparatus is often assumed to be primitive, potentially reminiscent of LECA (Cavalier-Smith 2010; Schneider et al. 2008). Apart from some exceptions including Tim17, most components of the mitochondrial protein import machinery generally present in other eukaryotes cannot be identified in euglenozoan

sequence data by homology-hit searches (Schneider et al. 2008; Singha et al. 2008). The most striking is the potential absence of Tom40 (Cavalier-Smith 2010; Pusnik et al. 2011; Schneider et al. 2008). Pusnik et al. (2011) have recently identified the protein translocase ATOM belonging to eukaryotic porin family in trypanosomes (Pusnik et al. 2011). ATOM is assumed to have the same function as Tom40 in other eukaryotes, and it has been suggested that this translocase is related to bacterial Omp85-like proteins (Pusnik et al. 2011). On the other hand, Hidden Markov Model (HMM)-based analysis of trypanosomatid ATOMs has recently revealed that ATOM is most likely highly derived Tom40 (Žárský et al. 2012). Our HMM search detected neither Tom40 nor ATOM in the transcriptomes of euglenids.

Since euglenozoan mitochondrial import machinery seems to be different from all organisms studied so far (Lithgow and Schneider 2010), one would expect that mitochondrial presequences would be also somehow specific. However, it has been recently shown that trypanosome mitochondria have the capacity to use the human CTS (Long et al. 2008). Moreover, except for highly variable sequence length and the presence of (M/L)RR motif in presequences, there seems to be nothing special about euglenozoan presequences in comparison to other organisms. The (M/L)RR logo has been previously generated, when N-termini of 24 *L. major* mitochondrial proteins precursors were included in the analysis (Uboldi et al. 2006). Euglenozoan presequences are generally rich in hydrophobic aa-s such as alanine, phenylalanine, leucine, and valine, and the basic arginine. (M/L)RR motif is present at the N-terminus (up to 10 aa) in most presequences and it is generally followed by a hydrophobic region, although in some longer presequences the RR motif is found more distantly from the N-terminus. In addition, when some *Bos taurus* and/or *S. cerevisiae* mitochondrial presequences were included in the analysis, they were all in the outputs of the MEME program (data not shown) suggesting that biochemical properties of euglenozoan presequences and motifs therein are not very different from the biochemical properties of presequences from other unrelated organisms. Nevertheless, the data presented here are consistent with the hypothesis that only the (M/L)RR motif and/or RR within two hydrophobic aa-s (mainly alanine, leucine, methionine or valine) or within longer hydrophobic regions is probably sufficient euglenozoan mitochondrial targeting signal irrespective of presequence length. Since not all presequences possess exactly (M/L)RR motif, but they all share at least short region with highly similar biochemical properties, another possibility is that only hydrophobic region rich in arginine (R) at the N-terminus of a preprotein can be sufficient signal for mitochondrial import in Euglenozoa. Nevertheless, it was also possible to identify 20 aa-long region with statistically significant biochemical similarities in 50 of 56 euglenozoan presequences (see Table 3 and Fig. 1e). Therefore, it is also possible that longer presequence regions with similar biochemical properties in long presequences can serve as N-terminal mitochondria-targeting signals. It has been also experimentally demonstrated that *T. brucei* CoxVI does not possess N-terminal cleavable presequence, and it uses internal signal sequence for targeting to mitochondria (Tasker et al. 2001). Taken altogether, although the majority of euglenozoan mitochondrial protein precursors likely requires only (M/L)RR motif and/or short hydrophobic region rich in arginine at the N-terminus to be targeted to mitochondria, longer motifs within longer presequences of some precursors and internal protein sequences can be also responsible for protein targeting to euglenozoan mitochondria.

Acknowledgments This work was supported by Scientific Grant Agency of the Slovak Ministry of Education and the Academy of Sciences (grants 1/0416/09 and 1/0393/09), Comenius University Grants (UK/54/2011), Czech Science Foundation grant P506/11/1320, and is the result of the project implementation: "The Improvement of Centre of excellence for exploitation of informational biomacromolecules in improvement of quality of life," ITMS 26240120027, supported by the Research & Development Operational Programme funded by the ERDF. This paper has been published in frame of the project "Strengthening research institutions at the University of Ostrava," CZ.1.07/2.3.00/30.0047, which is co-financed by the European Social Fund and the state budget of the Czech Republic. We thank Dr. Broňa Brejová (Department of Computer Science, Faculty of Mathematics, Physics, and Informatics, Comenius University, Bratislava, Slovakia), Dr. Tomáš Vinař (Department of Applied Informatics, Faculty of Mathematics, Physics, and Informatics, Comenius University, Bratislava, Slovakia), and Dr. Pavel Doležal and Vojtěch Žárský (both from the Department of Parasitology, Faculty of Science, Charles University in Prague) for help with the choice of appropriate bioinformatic programs to analyze the data.

References

Ahmadinejad N, Dagan T, Martin W (2007) Genome history in the symbiotic hybrid *Euglena gracilis*. Gene 402:35–39

Allen CA, Jackson AP, Rigden DJ, Willis AC, Ferguson SJ, Ginger ML (2008) Order within a mosaic distribution of mitochondrial c-type cytochrome biogenesis systems? FEBS J 275:2385–2402

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: proceedings of the second international conference on intelligent systems for molecular biology, AAAI Press, Menlo Park, pp 28–36

Bonen L (1993) *Trans*-splicing of pre-mRNA in plants, animals, and protists. FASEB J 7:40–46

Bourne PE, Addess KJ, Bluhm WF, Chen L, Deshpande N, Feng Z, Fleri W, Green R, Merino-Ott JC, Townsend-Merino W, Weissig H, Westbrook J, Berman HM (2004) The distribution and query systems of the RCSB protein data bank. Nucl Acids Res 32:D223–D225

Breglia SA, Yubuki N, Hoppenrath M, Leander BS (2010) Ultrastructure and molecular phylogenetic position of a novel

euglenozoan with extrusive episymbiotic bacteria: *Bihospites bacati* n. gen. et sp. (Symbiontida). BMC Microbiol 10:145
Bromley EV, Taylor MC, Wilkinson SR, Kelly JM (2004) The amino terminal domain of novel WD repeat protein from *Trypanosoma cruzi* contains a non-canonical mitochondrial targeting signal. Int J Parasitol 34:63–71
Callahan H, Litaker RW, Noga EJ (2002) Molecular taxonomy of the suborder Bodonina (Order Kinetoplastida), including the important fish parasite, *Ichthyobodo nacator*. J Eukaryot Microbiol 49:119–128
Cavalier-Smith T (2002) The phagotrophic origin of eukaryotes and the phylogenetic classification of Protozoa. Int J Syst Evol Microbiol 52:297–354
Cavalier-Smith T (2010) Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol Lett 6:342–345
Chan Y-F, Moestrup Ø, Chang J (2012) On *Keelungia pulex* nov. gen. et nov. sp., a heterotrophic euglenoid flagellate that lacks pellicular plates (Euglenophyceae, Euglenida). Eur J Protistol. http://dx.doi.org/10.1016/j.ejop.2012.04.003. Accessed 13 Aug 2012
Claros MG, Vincens P (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem 241:779–786
Cui J-Y, Mukai K, Saeki K, Matsubara H (1994) Molecular cloning and nucleotide sequences of cDNAs encoding subunits I, II and IX of *Euglena gracilis* mitochondrial complex III. J Biochem 115:98–107
Deschamps P, Lara E, Marande W, López-García P, Ekelund F, Moreira D (2011) Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. Mol Biol Evol 28:53–58
Desmond E, Brochier-Armanet C, Forterre P, Gribaldo S (2011) On the last common ancestor and early evolution of eukaryotes: Reconstructing the history of mitochondrial ribosomes. Res Microbiol 162:53–70
Di Giulio M (2007) The universal ancestor and the ancestors of Archaea and Bacteria were anaerobes whereas the ancestor of the Eukarya domain was an aerobe. J Evol Biol 20:543–548
Dooijes D, Chaves I, Kieft RA, Dirks-Mulder A, Martin W, Borst P (2000) Base J originally found in Kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. Nucl Acids Res 28:3017–3021
Durnford DG, Gray MW (2006) Analysis of *Euglena gracilis* plastid-targeted proteins reveals different classes of transit sequences. Eukaryot Cell 5:2079–2091
Dyková I, Fiala I, Lom J, Lukeš J (2003) *Perkinsiella* amoebae-like endosymbionts of *Neoparamoeba* spp., relatives of the kinetoplastid *Ichthyobodo*. Eur J Protistol 39:37–52
Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016
Ferreira VD, Rocchette I, Conforti V, Bench S, Feldman R, Levin MJ (2007) Gene expression patterns in *Euglena gracilis*: insight into the cellular response to environmental stress. Gene 389:136–145
Frantz C, Ebel C, Paulus F, Imbault P (2000) Characterization of trans-splicing in Euglenoids. Curr Genet 37:349–355
Frith MC, Saunders NFW, Kobe B, Bailey TL (2008) Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput Biol 4:e1000071
Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005). Protein identification and analysis tools on the ExPASy Server. In: Walker JM (ed) The proteomics protocols handbook, Humana Press, Totowa, pp 571–607
Gawryluk RMR, Gray MW (2009) A split and rearranged nuclear gene encoding the iron-sulfur subunit of mitochondrial succinate dehydrogenase in Euglenozoa. BMC Res Notes 2:16
Ginger ML, Fritz-Laylin LK, Fulton C, Cande WZ, Dawson SC (2010) Intermediary metabolism in protists: a sequence-based view of facultative anaerobic metabolism in evolutionary diverse eukaryotes. Protist 161:642–671
Hajduk SL, Harris ME, Pollard VW (1993) RNA editing in kinetoplastid mitochondria. FASEB J 7:54–63
Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci USA 106:3859–3864
Häusler T, Stierhof YD, Blattner J, Clayton C (1997) Conservation of mitochondrial targeting sequence function in mitochondrial and hydrogenosomal proteins from the early-branching eukaryotes *Crithidia*, *Trypanosoma* and *Trichomonas*. Eur J Cell Biol 73:240–251
Kořený L, Oborník M (2011) Sequence evidence for the presence of two tetrapyrrole pathways in *Euglena gracilis*. Genome Bio Evol 3:359–364
Leander BS (2004) Did trypanosomatid parasites have photosynthetic ancestors? Trends Microbiol 12:251–258
Leander BS, Triemer RE, Farmer MA (2001) Character evolution in heterotrophic euglenids. Eur J Protistol 37:337–356
Liang XH, Haritan A, Uliel S, Michaeli S (2003) *Trans* and *cis* splicing in trypanosomatids: mechanism, factors, and regulation. Eukaryot Cell 2:830–840
Likic VA, Doležal P, Celik N, Dagley M, Lithgow T (2010) Using hidden markov models to discover new protein transport machineries. Methods Mol Biol 619:271–284
Linton EW, Karnkowska-Ishikawa A, Kim JI, Shin W, Bennett MS, Kwiatowski J, Zakryś B, Triemer RE (2010) Reconstructing euglenoid evolutionary relationships using three genes: nuclear SSU and LSU, and chloroplast SSU rDNA sequences and the description of *Euglenaria* gen. nov. (Euglenophyta). Protist 161:603–619
Lithgow T, Schneider A (2010) Evolution of macromolecular import pathways in mitochondria, hydrogenosomes and mitosomes. Phil Trans R Soc B 365:799–817
Long S, Jirků M, Ayala FJ, Lukeš J (2008) Mitochondrial localization of human frataxin is necessary but processing is not for rescuing frataxin deficiency in *Trypanosoma brucei*. Proc Natl Acad Sci USA 105:1373–13468
Marande W, Burger G (2007) Mitochondrial DNA as a genomic jigsaw puzzle. Science 318:415
Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. Nature 392:37–41
Maslov DA, Zíková A, Kyselová I, Lukeš J (2002) A putative novel nuclear-encoded subunit of cytochrome *c* oxidase complex in trypanosomatids. Mol Biochem Parasitol 125:113–225
Moreira D, López-García P, Vickerman K (2004) An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the Class kinetoplastea. Int J Syst Evol Microbiol 54:1861–1875
Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10:1–6
Nozaki H, Ohta N, Matsuzaki M, Misumi O, Kuroiwa T (2003) Phylogeny of plastids based on cladistic analysis of gene loss inferred from complete plastid genome sequences. J Mol Evol 57:377–382
Priest JW, Hajduk SL (1996) In vitro import of the Rieske iron-sulfur protein by trypanosome mitochondria. J Biol Chem 271:20060–20069
Priest JW, Hajduk SL (2003) *Trypanosoma brucei* cytochrome c1 is imported into mitochondria along an unusual pathway. J Biol Chem 287:15084–15094

Priest JW, Wood ZA, Hajduk SL (1993) Cytochromes c1 of kinetoplastid protozoa lack mitochondrial targeting presequences. Biochim Biophys Acta 1144:229–231

Pusnik M, Schmidt O, Perry AJ, Oeljeklaus S, Niemann M, Warcheid B, Lithgow T, Meisinger C, Schneider A (2011) Mitochondrial preprotein translocase of trypanosomatids has a bacterial origin. Curr Biol 21:1738–1743

Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ (2007) The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. Mol Biol Evol 24:54–62

Schneider A, Bursa D, Lithgow T (2008) The direct route: a simplified pathway for protein import into the mitochondrion of trypanosomes. Trends Cell Biol 18:12–18

Simpson AGB (1997) The identity and composition of the Euglenozoa. Arch Protistenkd 148:318–328

Simpson AGB, Roger AJ (2004) Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. Mol Phylogenet Evol 30:201–212

Simpson L, Thiemann OH, Savill NJ, Alfonzo JD, Maslov DA (2000) Evolution of RNA editing in trypanosome mitochondria. Proc Natl Acad Sci USA 97:6986–6993

Simpson AGB, Lukeš J, Roger AJ (2002) The evolutionary history of kinetoplastids and their kinetoplasts. Mol Biol Evol 19: 2071–2083

Simpson AGB, Gill EE, Callahan HA, Litaker RW, Roger AJ (2004) Early evolution within kinetoplastids (Euglenozoa), and the late emergence of trypanosomatids. Protist 155:407–422

Simpson AGB, Stevens JR, Lukeš J (2006) The evolution of kinetoplastid flagellates. Trends Parasitol 22:168–174

Singha UK, Paprah E, Williams R, Saha L, Chaudhuri M (2008) Characterization of the mitochondrial inner protein translocator Tim17 from *Trypanosoma brucei*. Mol Biochem Parasitol 159:30–43

Söding J (2005) Protein homology detection by HMM–HMM comparison. Bioinformatics 21:951–960

Spencer DF, Gray MW (2011) Ribosomal RNA genes in *Euglena gracilis* mitochondrial DNA: fragmented genes in a seemingly fragmented genome. Mol Genet Genomics 285:19–31

Stuart K, Panigrahi AK (2002) RNA editing: complexity and complications. Mol Microbiol 45:591–596

Tasker M, Timms M, Hendriks E, Matthews K (2001) Cytochrome oxidase subunit VI of *Trypanosoma brucei* is imported without a cleaved presequence and is developmentally regulated at both RNA and protein levels. Mol Microbiol 39:272–285

Triemer RE, Farmer MA (1991) An ultrastructural comparison of the mitotic apparatus, feeding apparatus, flagellar apparatus and cytoskeleton in euglenoids and kinetoplastids. Protoplasma 164:91–104

Turmel M, Gagnon M-C, O'Kelly CJ, Lemieux C (2009) The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol Biol Evol 26:631–648

Uboldi AD, Lueder FB, Walsh P, Spurck T, McFadden GI, Curtis J, Likic VA, Perugini MA, Barson M, Lithgow T, Handman E (2006) A mitochondrial protein affects cell morphology, mitochondrial segregation and virulence in *Leishmania*. Int J Parasitol 36:1499–1514

Vesteg M, Krajčovič J (2008) Origin of eukaryotic cells as a symbiosis of parasitic α-proteobacteria in the periplasm of two-membrane-bounded sexual pre-karyotes. Commun Integr Biol 1:104–113

Vesteg M, Krajčovič J (2011) The falsifiability of the models for the origin of eukaryotes. Curr Genet 57:367–390

Vesteg M, Vacula R, Steiner JM, Mateášiková B, Löffelhardt W, Brejová B, Krajčovič J (2010) A possible role for short introns in the acquisition of stroma-targeting peptides in the flagellate *Euglena gracilis*. DNA Res 17:223–231

Vlček Č, Marande W, Teijeiro S, Lukeš J, Burger G (2011) Systematically fragmented genes in a multipartite mitochondrial genome. Nucl Acids Res 39:979–988

von der Heyden S, Chao EE, Vickerman K, Cavalier-Smith T (2004) Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of Euglenozoa. J Eukaryot Microbiol 51:402–416

Williams S, Saha L, Singha UK, Chaudhuri M (2008) *Trypanosoma brucei*: differential requirement of membrane potential for import of proteins into mitochondria in two developmental stages. Exp Parasitol 118:420–433

Yamaguchi A, Yubuki N, Leander BS (2012) Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy: description of *Rapaza viridis* n. gen. et sp. (Euglenozoa, Euglenida). BMC Evol Biol 12:29

Yubuki N, Edgcomb VP, Bernhardt JM, Leander BS (2009) Ultrastructure and molecular phylogeny of *Calkinsia aureus*: cellular identity of a novel clade of deep-sea euglenozoans with epibiotic bacteria. BMC Microbiol 9:16

Žárský V, Tachezy J, Doležal P (2012) Tom40 is likely common to all mitochondria. Curr Biol 22:R479–R481