Available online at www.sciencedirect.com
ScienceDirect
Gene 402 (2007) 35–39
GENE
www.elsevier.com/locate/gene

Genome history in the symbiotic hybrid *Euglena gracilis*

Nahal Ahmadinejad, Tal Dagan *, William Martin

Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany

Received 19 April 2007; received in revised form 16 July 2007; accepted 16 July 2007
Available online 2 August 2007
Received by G. Theissen

Abstract

Euglena gracilis has a chimeric gene collection in which some genes were inherited from its heterotrophic host and others were acquired from a photoautotrophic endosymbiont during secondary endosymbiosis. The evolutionary reconstruction of such a hybrid genome poses a challenge for standard phylogenetic tools that produce bifurcating trees because genome evolution by endosymbiotic gene transfer is a non tree-like process. We sequenced 2770 ESTs from *E. gracilis*, of which 841 have homologues in a sample of other eukaryotes. Most of these homologues are found in all of the eukaryotes in our sample, but 117 of them are specific to photoautotrophic eukaryotes. A phylogenetic tree fails to account for this observation but the distribution of homologues and a phylogenetic network clearly show the common origin of *E. gracilis* from both kinetoplastid and photoautotrophic ancestors.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Endosymbiotic transfer; Genome evolution; Phylogenetic networks

1. Introduction

Most photoautotrophic eukaryotes acquired their photosynthetic lifestyle from a cyanobacterial endosymbiont (McFadden, 2001). The genome of the endosymbiont itself has since then been reduced and the endosymbiont evolved into a DNA possessing organelle – a primary plastid – as found in green and red algae, land plants and glaucocystophytes (Adl et al., 2005). Several independent eukaryotic lineages have acquired their photosynthetic lifestyle from a secondary endosymbiont — a eukaryotic alga that became engulfed by another eukaryote (Gibbs, 1978; Stoebe and Maier, 2002). Both primary and secondary endosymbiosis were accompanied by endosymbiotic gene transfer (EGT) — the relocation of genes from the organelle to the chromosomes of the host (Martin et al., 1993; Archibald et al., 2003; Timmis et al., 2004). Estimations for the frequency of EGT during the primary endosymbiosis range between 18% of the *Arabidopsis thaliana* genome (Martin et al., 2002) and 11% of the *Cyanophora* genome (Reyes-Prieto et al., 2006).

Euglena gracilis is well suited for the study of endosymbiosis and endosymbiotic gene transfer because its plastid was acquired by a secondary endosymbiosis, but it includes no remains of the endosymbiotic nucleus (McFadden, 2001). Moreover, *E. gracilis* shares a common ancestor with the Kinetoplastida (Adl et al., 2005), none of which seem to have experienced a secondary endosymbiosis (Rogers et al., 2006). Hence, by the sequence similarity criterion, the genome of *E. gracilis* is expected to be a hybrid composed of four main gene classes: (i) *Euglena*-specific genes, (ii) Kinetoplastida-specific genes, (iii) eukaryotic genes that are spread in other eukaryotes, and (iv) genes acquired during the secondary endosymbiosis. Clear candidates for the latter group are genes that have homologues only in photoautotrophic eukaryotes. In some cases the acquired gene replaced an orthologous gene within the host (Henze et al., 1995). Such genes have homologues in both photoautotrophic and heterotrophic eukaryotes, but are closely related to the former. Because of these distinct gene origins, a single bifurcating tree cannot accurately describe the chimeric gene collection of *E. gracilis* because genome evolution by endosymbiotic gene transfer is a non tree-like process. Here, we describe the genome composition of *E. gracilis* as estimated

Abbreviations: BBH, Best blast hit; EGT, Endosymbiotic gene transfer; EST, Expressed Sequence Tag; ML, Maximum likelihood; NJ, Neighbor joining.

* Corresponding author. Tel.: +49 211 8112736; fax: +49 211 8113554.
E-mail address: tal.dagan@uni-duesseldorf.de (T. Dagan).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.07.023

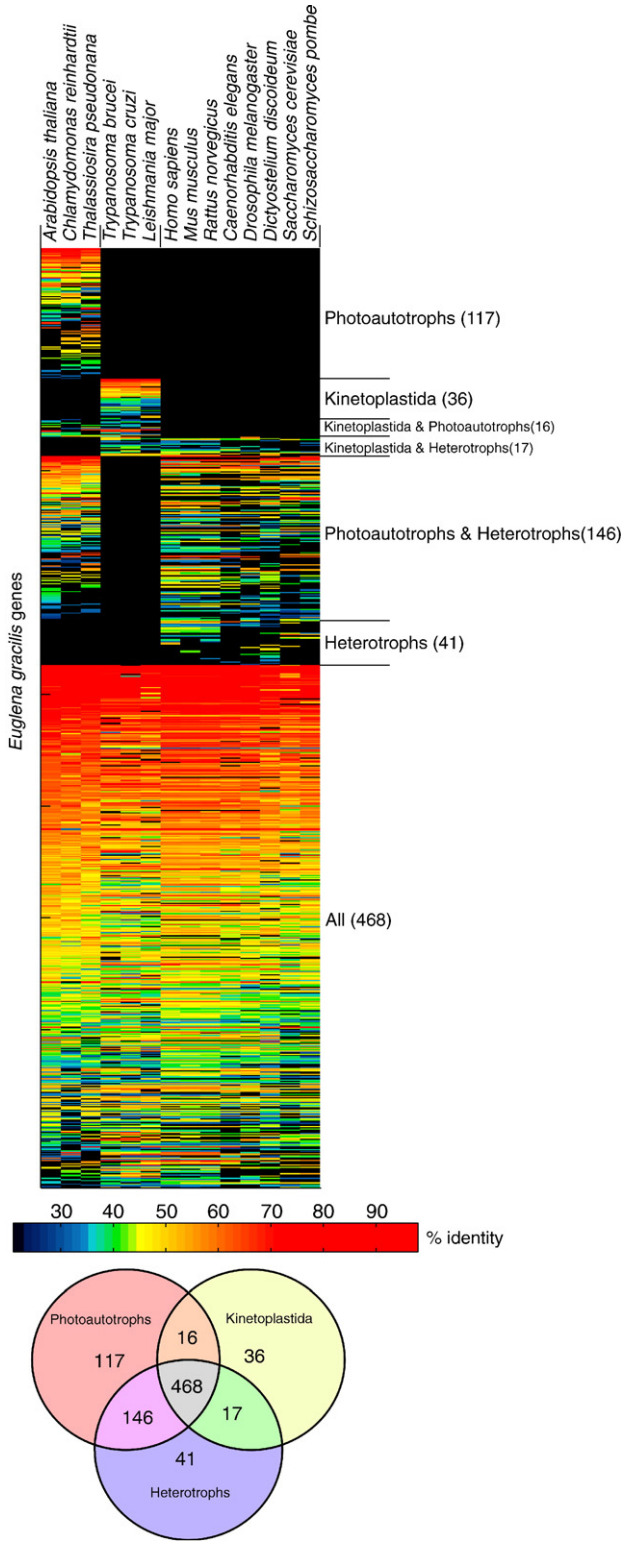

Fig. 1. The distribution of homologues to *E. gracilis* genes in genomes of photoautotrophic eukaryotes, Kinetoplastida, and heterotrophic eukaryotes. Columns indicate *Euglena gracilis* proteins, rows indicate reference genomes, amino acid identity in pairwise protein sequence comparisons is color-coded as shown in the scale. A Venn diagram at the bottom presents the distribution of common genes for *E. gracilis* and the reference genomes.

from random cDNA sequences while investigating methods for reconstructing the genomic history of symbiotic hybrids.

2. Materials and methods

From *E. gracilis* cDNA libraries prepared from cells grown under aerobic and anaerobic conditions (Rotte et al., 2001), 10,793 clones (5450 from aerobically and 5343 from anaerobically cultured cells) were sequenced. That fell into 2667 unique sequences (accessions EL579575–EL582344 in dbEST). The ESTs were BLASTed (Altschul et al., 1997) against 14 eukaryotic genomes (Table S1) using BLASTX. Reading frames were determined by BLASTX alignment. Forty-one ESTs harbored conflicting reading frames and were corrected manually. ESTs were translated into proteins using the EMBOSS package (Rice et al., 2000) and classified into functional categories according to their BBH in the Swiss-Prot database (Boeckmann et al., 2003).

In the sequence similarity approach, the nearest neighbor was inferred as the best BLASTX hit (BBH) of the gene in question, where only hits of *e*-value $\leq 10^{-10}$ and $\geq 25\%$ identities were considered. Each *E. gracilis* protein with BBHs in at least 4 taxa (615 genes) was aligned with its BBHs using CLUSTALW (Thompson et al., 1994) and all gap-containing sites were removed. Phylogenetic trees by maximum-likelihood approach were reconstructed with FastML (Pupko et al., 2000). This left us with 587 ML gene-trees. The same alignments were used to calculate protein distances with PROTDIST (PHYLIP; Felsenstein, 2005), and neighbor-joining trees with NEIGHBOR (PHYLIP; Felsenstein, 2005). To identify the nearest neighbor by the phylogenetic approach we extracted the tree splits using CONSENSE (PHYLIP; Felsenstein, 2005). The nearest neighbor was then classified according to the members of the smallest clade that includes *E. gracilis*.

A consensus phylogenetic tree was reconstructed from the gene-trees of the 259 globally distributed genes using CONSENSE (PHYLIP; Felsenstein, 2005). The 259 alignments were concatenated to form a 36,570 amino acids long alignment that was used to reconstruct a neighbor-joining tree (Saitou and Nei,

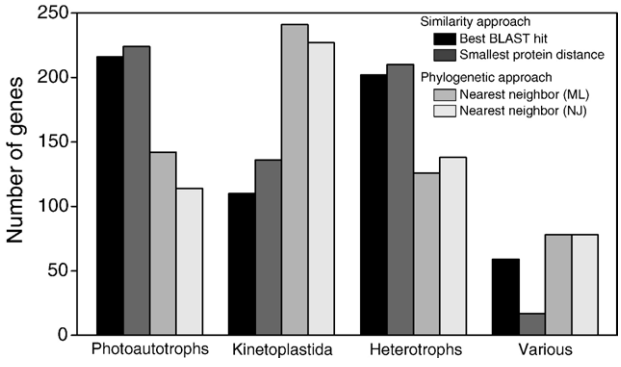

Fig. 2. A distribution of the nearest neighbors to *E. gracilis* genes across the three tested groups by the similarity and phylogenetic approaches. When the nearest neighbor could not be determined the gene was classified into the "Various" group.

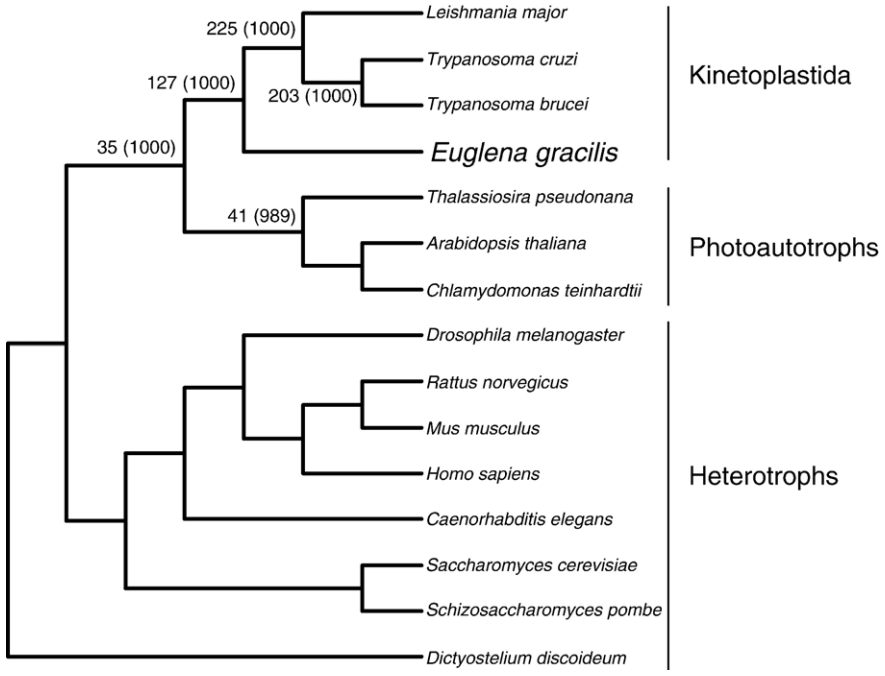

Fig. 3. A phylogenetic tree of 259 globally distributed genes. The number of gene-trees concordant with the topology are indicated on the nodes (left) as well as the number of bootstrap replications supporting the node (right).

1987) with CLUSTALW (Thompson et al., 1994). For the phylogenetic network pairwise distances were computed with LDDist (Thollesson, 2004), splits were calculated with NeighborNet (NNet; Bryant and Moulton, 2004), and represented as a planar graph with SplitsTree v4.2 (Huson and Bryant, 2006).

3. Results and discussion

We sequenced a sample of 2770 *E. gracilis* ESTs, of which 841 sequences had homologues in a sample of 14 other eukaryotic genomes. One approach to examine the genomic history of symbiotic hybrids is to examine patterns of shared

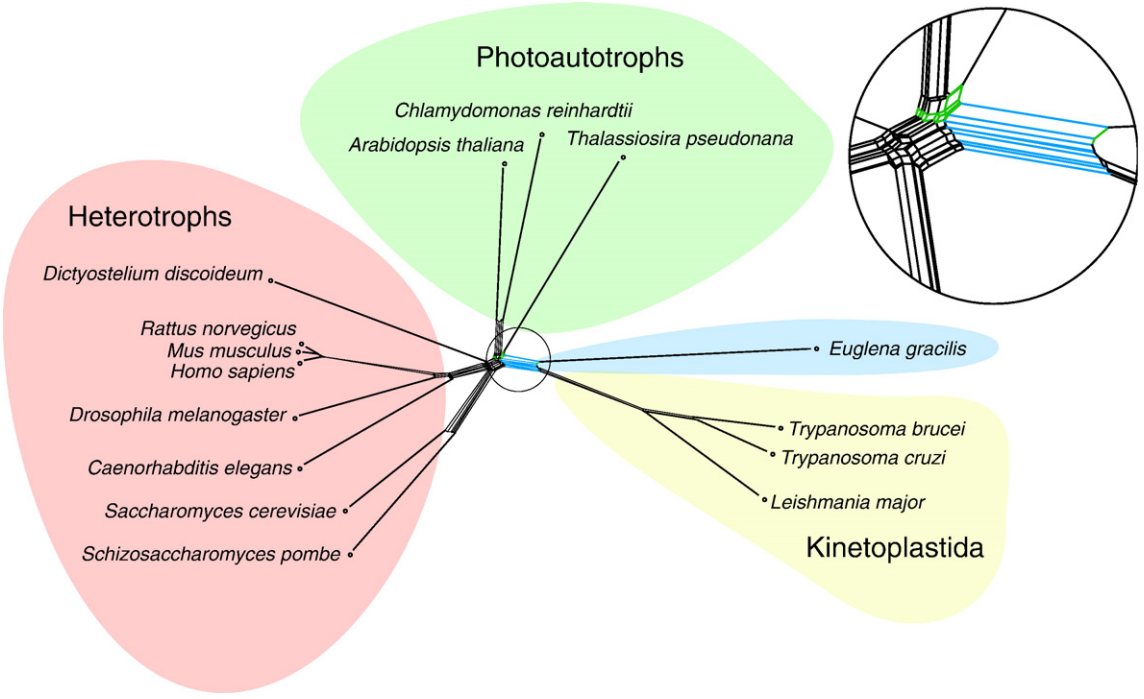

Fig. 4. A phylogenetic network reconstructed for the concatenated alignments of 259 globally distributed genes. The relevant splits are enlarged on the right.

genes. We classified the 841 genes into the following gene classes: genes that have homologues in Kinetoplastida, homologues in heterotrophic eukaryotes (animals and fungi), homologues in photoautotrophic eukaryotes (a plant, a green alga and a diatom) and all of the combinations thereof (Fig. 1). About 14% of the genes have homologues only in photoautotrophic eukaryotes; of these 31 (25%) genes are involved in photosynthesis. This proportion is significantly above the background similarities as measured by the share of photosynthetic genes in the rest of the sample (1%; $p<0.05$ using χ^2 test). Only 36 (4%) of the *E. gracilis* genes are Kinetoplastida-specific, most of them are ORFs with unknown function (Fig. 1).

3.1. Nearest neighbor analysis

Another method for classifying the genes into the above three groups is by the nearest neighbor method. In this approach, the gene is classified into the group of its closest related (i.e., most similar) sequence. We employed the nearest neighbor method by two approaches, first by using only sequence similarity — the nearest neighbor is the homologue with the highest amino acid identity. Second we reconstructed phylogenetic trees in which the nearest neighbor is the taxon or taxa that branch with the *E. gracilis* gene. In the similarity approach we tested the identity of the best blast hit (BBH) of the *E. gracilis* gene and in addition the most similar protein in a protein distance matrix that was calculated from the multiple sequence alignment of all homologues found for the *E. gracilis* gene. Both similarity approaches resulted in similar distribution of nearest neighbors where Kinetoplastidan homologues are the minority, while the proportion of photoautotrophic and heterotrophic nearest neighbors is similar (Fig. 2). Hence in this case, the BBH results, which may not always imply phylogenetic relationship (Koski and Golding, 2001), are supported by similar results of the protein distances. By contrast, inference of the nearest neighbors using the phylogenetic approach with both ML and NJ trees yields different distributions where most of the nearest neighbors are in the Kinetoplastida and the remainder are distributed among photoautotrophic and heterotrophic eukaryotes (Fig. 2), whereby the difference between the similarity and phylogenetic approaches is independent of the tree reconstruction method (ML vs. NJ) in the present case.

An example of discrepancy between nearest neighbor relationships as assessed by phylogenetic and similarity approaches is found in the case of the *E. gracilis* β-tubulin protein. Homologues for the gene were found in all of the genomes in our analysis, and their multiple sequence alignment is highly reliable with 94% of the aligned columns and 99% of the pairs resulting in identical head and tail alignments by the HoT method (Landan and Graur, 2007). The distances of *E. gracilis* sequence from the photoautotrophs are smaller than the distance from the Kinetoplastida, yet *E. gracilis* branches with kinetoplastids in the β-tubulin phylogenetic tree (Supplementary Figure 1). Accordingly, we have presented the results of both approaches here (Fig. 2).

3.2. Phylogeny of globally distributed genes

A phylogenetic tree reconstructed using the concatenated alignments of 259 globally distributed genes with neighbor-joining approach (Saitou and Nei, 1987) shows a clear grouping of *E. gracilis* with the Kinetoplastida (Fig. 3). However, using a concatenated alignment averages the signal of all genes, even though some genes may have fundamentally different gene-trees (Huson and Bryant, 2006). Therefore we reconstructed the phylogenetic tree of each gene independently using maximum-likelihood approach. Only 127 (49%) gene-trees have a topology in which *E. gracilis* branches with the Kinetoplastida (Fig. 3), whereas in 57 (22%) of the gene-trees *E. gracilis* branches with one or more photoautotrophic eukaryotes. For example, in the phylogenetic gene-trees of chloroplast enolase and GAPDH, *E. gracilis* and the photoautotrophs form a monophyletic group. These two cases are documented as orthologous replacements by endosymbiotic acquired genes (Hannaert et al., 2000; Henze et al., 1995). In a phylogenetic network reconstructed for the concatenated alignments of the globally distributed genes *E. gracilis* shares a strong split with the Kinetoplastida, but also a split with the photoautotrophic eukaryotes (Fig. 4). The weight of the Kinetoplastida split (0.03) is much larger than the photoautotroph split (0.005), and this reflects the circumstance that in most of the globally distributed genes the Kinetoplastida are the closest relatives of *E. gracilis*, but in some of them the gene from the photoautotrophs is the closest relative.

3.3. EGT during secondary endosymbiosis

Estimates for the proportion of EGTs following secondary endosymbiosis are lacking. In our sample, *E. gracilis* shares most of its genes with both photoautotrophic and heterotrophic eukaryotes. About 13% (by the similarity measure) or 10% (by the phylogenetic measure) of *E. gracilis* genes had their nearest neighbors in the genomes of photoautotrophic eukaryotes, but only 4.5% of the genes had homologues exclusively in that group. We consider the latter proportion as a conservative estimate for the percent of genes that were acquired by *E. gracilis* during the secondary endosymbiosis. An EST-based survey focusing on 78 secondary plastid-targeted genes in the alga *Bigelowiella natans* identified 49 likely secondary EGT events using gene-trees (Archibald et al., 2003), which would represent 1.2% of the *B. natans* genes in that sample, but that value was estimated only using proteins that were predicted to plastid-targeted.

4. Conclusions

The genome of *E. gracilis* is a hybrid of photoautotrophic and heterotrophic genomes as the nearest neighbors of its genes attest. However, no phylogenetic tree can illustrate this evolutionary history because phylogenetic trees are designed to reflect only scenarios of vertical evolution with descent from a single common ancestor (Dagan and Martin, 2006), whereas the *E. gracilis* genome has two distinct ancestors. The

phylogenetic tree reconstructed for the globally distributed genes in our sample results in a monophyletic clade of *Euglena* and Kinetoplastida that is supported by high bootstrap proportions, and the majority rule of consensus gene-trees. But this topology shows only half of the evolutionary history of *E. gracilis*; the other half – the contribution of the endosymbiont to the host genome – is not represented by the bifurcating tree. The alternative narrative is being told by the distribution of the homologues to *E. gracilis* genes over the eukaryotic domain (Fig. 1) or the nearest neighbors by both approaches (Fig. 2). But the most telling single-graph analysis would appear to be the phylogenetic network that is able to uncovers conflicting signals in molecular data (Fitch, 1997; Bryant and Moulton, 2004; Holland et al., 2004). Our network uncovers the conflicting phylogenetic signals linking *E. gracilis* with both the Kinetoplastida and the photoautotrophs.

Acknowledgements

We thank Gabriel Gelius-Dietrich for his help in analyzing the data and Katrin Henze and Toni Gabaldón for helpful feedback.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.07.023.

References

Adl, S.M., et al., 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J. Eukaryot. Microbiol. 52, 399–451.

Altschul, S.F., et al., 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Archibald, J.M., Rogers, M.B., Toop, M., Ishida, K., Keeling, P.J., 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc. Natl. Acad. Sci. U. S. A. 100, 7678–7683.

Boeckmann, B., et al., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370.

Bryant, D., Moulton, V., 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21, 255–265.

Dagan, T., Martin, W., 2006. The tree of one percent. Genome. Biol. 7, 118.

Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package). Department of Genome Sciences, University of Washington, Seattle.

Fitch, W., 1997. Networks and viral evolution. J. Mol. Evol. 44, S65–S75.

Gibbs, S.P., 1978. The chloroplast of *Euglena* may have evolved from symbiotic green algae. Can. J. Bot. 56, 2883–2889.

Hannaert, V., et al., 2000. Enolase from *Trypanosoma brucei*, from the amitochondriate protist *Mastigamoeba balamuthi*, and from the chloroplast and cytosol of *Euglena gracilis*: pieces in the evolutionary puzzle of the eukaryotic glycolytic pathway. Mol. Biol. Evol. 17, 989–1000.

Henze, K., Badr, A., Wettern, M., Cerff, R., Martin, W., 1995. A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. Proc. Natl. Acad. Sci. U. S. A. 92, 9122–9126.

Holland, B.R., Huber, K.T., Moulton, V., Lockhart, P.J., 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol. Biol. Evol. 21, 1459–1461.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Koski, L.B., Golding, G.B., 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52, 540–542.

Landan, G., Graur, D., 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol. Biol. Evol. 24, 1380–1383.

Martin, W., Brinkmann, H., Savonna, C., Cerff, R., 1993. Evidence for a chimeric nature of nuclear genomes — eubacterial origin of eukaryotic glycerinaldehyde-3-phosphate dehydrogenase genes. Proc. Natl. Acad. Sci. U. S. A. 90, 8692–8696.

Martin, W., et al., 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. U. S. A. 99, 12246–12251.

McFadden, G.I., 2001. Primary and secondary endosymbiosis and the origin of plastids. J. Phycol. 37, 951–959.

Pupko, T., Pe'er, I., Shamir, R., Graur, D., 2000. A fast algorithm for joint reconstruction of ancestral amino-acid sequences. Mol. Biol. Evol. 17, 890–896.

Reyes-Prieto, A., Hackett, J.D., Soares, M.B., Bonaldo, M.F., Bhattacharya, D., 2006. Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. Curr. Biol. 16, 2320–2325.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.

Rogers, M.B., Gilson, P.R., Su, V., McFadden, G.I., Keeling, P.J., 2006. 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.

Rotte, C., Stejskal, F., Zhu, G., Keithly, J.S., Martin, W., 2001. Pyruvate: NADP+oxidoreductase from the mitochondrion of *Euglena gracilis* and from the apicomplexan *Cryptosporidium parvum*: a biochemical relic linking pyruvate metabolism in mitochondriate and amitochondriate protists. Mol. Biol. Evol. 18, 710–720.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstruction of phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Stoebe, B., Maier, U.G., 2002. One, two, three: nature's toolbox for building plastids. Protoplasma 219, 123–130.

Thollesson, M., 2004. LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics 20, 416–418.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., Martin, W., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev., Genet. 5, 123–135.