Tryptophan Biosynthesis in Stramenopiles: Eukaryotic Winners in the Diatom Complex Chloroplast

ARTICLE *in* JOURNAL OF MOLECULAR EVOLUTION · JANUARY 2007

Impact Factor: 1.68 · DOI: 10.1007/s00239-007-9022-z

J Mol Evol
DOI 10.1007/s00239-007-9022-z

Tryptophan Biosynthesis in Stramenopiles: Eukaryotic Winners in the Diatom Complex Chloroplast

Kateřina Jiroutová · Aleš Horák · Chris Bowler · Miroslav Oborník

Received: 5 January 2007 / Accepted: 2 July 2007

Abstract Tryptophan is an essential amino acid that, in eukaryotes, is synthesized either in the plastids of photoautotrophs or in the cytosol of fungi and oomycetes. Here we present an in silico analysis of the tryptophan biosynthetic pathway in stramenopiles, based on analysis of the genomes of the oomycetes *Phytophthora sojae* and *P. ramorum* and the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. Although the complete pathway is putatively located in the complex chloroplast of diatoms, only one of the involved enzymes, indole-3-glycerol phosphate synthase (InGPS), displays a possible cyanobacterial origin. On the other hand, in *P. tricornutum* this gene is fused with the cyanobacteria-derived hypothetical protein COG4398. Anthranilate synthase is also fused in diatoms. This fusion gene is almost certainly of bacterial origin, although the particular source of the gene cannot be resolved. All other diatom enzymes originate from the nucleus of the primary host (red alga) or secondary host (ancestor of chromalveolates). The entire pathway is of eukaryotic origin and cytosolic localization in oomycetes; however, one of the enzymes, anthranilate phosphoribosyl transferase, was likely transferred to the oomycete nucleus from the red algal nucleus during secondary endosymbiosis. This suggests possible retention of the complex plastid in the ancestor of stramenopiles and later loss of this organelle in oomycetes.

Keywords Tryptophan synthesis · Mosaic origin · Diatom · Oomycetes

Introduction

Diatoms (stramenopiles; Bacillariophyta) are unicellular photosynthetic eukaryotes that are responsible for about 20% of global carbon fixation. They contain a complex chloroplast, which is thought to originate via secondary endosymbiosis (Armbrust et al. 2004). In this process, which is proposed to have occurred about 1.3 billion years ago, the heterotrophic eukaryotic ancestor engulfed a photosynthetic red alga (Yoon et al. 2004). The algal endosymbiont subsequently developed into a variety of complex chloroplasts surrounded by more than two membranes. As proposed by Thomas Cavalier-Smith (1999, 2002), secondary endosymbiosis is a very complex process, which has involved transfer of many genes from the engulfed red alga to the secondary host nucleus, accompanied by a reduction of the red algal cellular structures and organelles. Consequently, it has probably happened only once in evolution and provided a photosynthetic

Electronic supplementary material The online version of this article (doi:10.1007/s00239-007-9022-z) contains supplementary material, which is available to authorized users.

K. Jiroutová · A. Horák · M. Oborník
Institute of Parasitology, Biology Centre of the Academy of Sciences of the Czech Republic, Branišovská 31, 37005 České Budějovice, Czech Republic

K. Jiroutová · A. Horák · M. Oborník (✉)
University of South Bohemia, Faculty of Biological Sciences, Branišovská 31, 37005 České Budějovice, Czech Republic
e-mail: obornik@paru.cas.cz

C. Bowler
CNRS UMR8186, Ecole Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France

C. Bowler
Cell Signalling Laboratory, Stazione Zoologica, Villa Comunale I, 80121 Naples, Italy

organelle for the whole group of Chromalveolata, organisms that had been heterotrophs and predators before this event (Cavalier-Smith 1999, 2002). However, other explanations have also been proposed to clarify the outstanding diversity of the red complex chloroplasts. These scenarios involve either the concept of separated endosymbioses for each group (Falkowski et al. 2004) or numerous tertiary endosymbiotic events (Bodyl 2005).

Tryptophan is an essential aromatic amino acid which must either be synthesized by the cell or be obtained from the environment. Within the eukaryotes, tryptophan is synthesized either in the chloroplasts, as in photoautotrophs, or in the cytosol, as in fungi and oomycetes. All other eukaryotic heterotrophs obtain aromatic amino acids from their food. Tryptophan is synthesized in seven steps from chorismate and five enzymes are involved in the pathway: anthranilate synthase (AS; α+β subunits in prokaryotes or components I and II in eukaryotes), anthranilate phosphoribosyl transferase (APRT), phosphoribosyl anthranilate isomerase (PRAI), indole-3-glycerol-phosphate synthase (InGPS), and tryptophan synthase (TS; α+β subunits), in order of their action in tryptophan biosynthesis (Crawford 1975) (see Fig. 1). The synthesis starts with an enolpyruvyl elimination accompanied by a glutamine amido transfer catalyzed by AS. As with other glutamine amidotransferases, under certain conditions, ammonia can replace glutamine in the reaction, giving anthranilate and pyruvate as products. This version of the reaction can usually be catalyzed by a single α subunit of the enzyme and is referred to as the chorismate amination reaction (Green and Nichols 1991). The addition of the phosphoribosyl moiety of phosphoribosyl pyrophosphate to anthranilate, which is catalyzed by APRT, is the second step in the pathway. The third step is an arrangement of phosphoribosyl mediated by anthranilate isomerase (PRAI) and is followed by decarboxylation and ring closure catalyzed by InGPS. Removal of the glycerol phosphate side chain from indole glycerol phosphate and its replacement by the alanyl part of L-serine is catalyzed by tryptophan synthase subunits α and β, respectively, and represents the final step in the biosynthetic pathway (Creigton and Yanofsky 1970) (for details see Fig. 1).

Although prokaryotes and eukaryotes generally use the same reactions to synthesize tryptophan, the genes involved differ in their arrangement and, in particular, in the occurrence of various gene fusions (Hütter 1986) (see Table 1 for details). In prokaryotes, the involved enzymes are encoded by seven genes: trpE and trpG encode the two subunits of AS, trpD encodes APRT, trpF PRAI, trpC InGPS, and trpB and trpA encode the two subunits of TS, respectively. However, various combinations of gene fusions have been found in prokaryotes (Yanofsky et al. 1981; Hütter 1986; Braus 1991). For instance, α and β subunits of AS are fused in some of the α-proteobacteria (Bae and Crawford 1990) and cyanobacteria, whereas cyanobacteria possess both the fusion gene and at least one of the individually encoded subunits of ASs. Genes encoding enzymes of tryptophan biosynthesis (trp genes) are usually arranged on the bacterial chromosome as gene clusters (Crawford 1989). Such clusters are, as shown in *Escherichia coli* and *Lactococcus lactis*, transcribed as a single transcript and thus form operons, which allow simultaneous regulation of the expression of all their

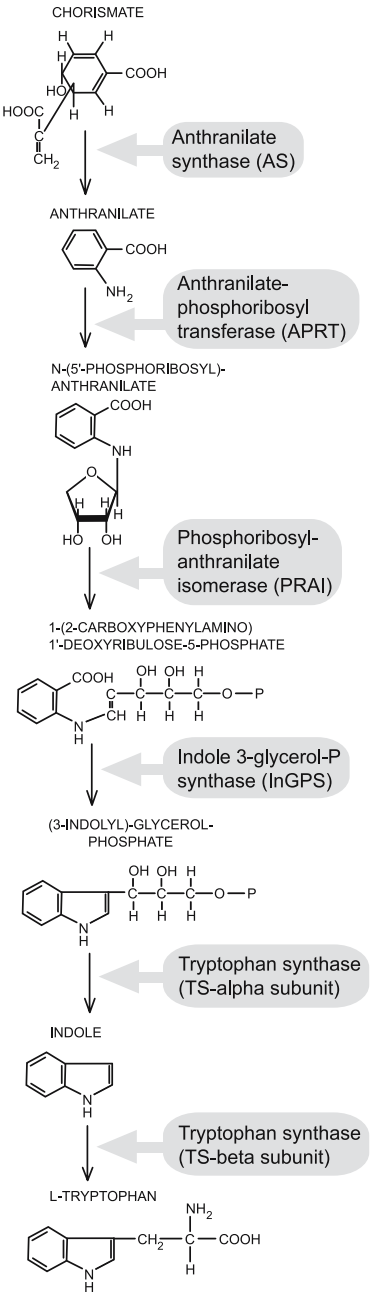


Fig. 1 Tryptophan biosynthesis

Table 1 Occurrence of genes involved in the synthesis of tryptophan in selected prokaryotes and eukaryotes

eukaryotes	**diatoms** (*Phaeodactylum tricornutum*)	pt **ASα + ASβ** 2,877 bp	pt **APRT** 1,337 bp	pt **PRAI + InGPS** 2,470 bp	pt **InGPS** 1,187 bp pt **InGPS** 1,209 bp pt **COG 4398 + InGPS** 2,979 bp	pt **TS α + TS β** 2,429 bp	**TS β** 1,939 bp only in *P. tricornutum*	
	plants (*Oryza sativa*)	pt **AS I** 1,821 bp M **AS I** 1,734 bp	pt **AS II** 1,143 bp	pt **APRT** 1,230 bp	pt **PRAI** 837 bp	pt **InGPS** 1,146 bp	pt **TS α** 1,248 bp	**TS β** 1,413 bp **TS β** 1,416 bp
	rhodophytes (*Cyanidioschyzon merolae*)	pt **AS I** 1,941 bp	pt **AS II** 561 bp	pt **APRT** 1,395 bp	pt **PRAI** 849 bp	pt **InGPS** 1,137 bp M **InGPS** 960 bp pt **InGPS** 1167 bp	pt **TS α** 729 bp **plastid encoded**	**TS β** 1,530
prokaryotes	**oomycetes** (*Phytophthora ramorum*)	**AS I** 1,524 bp	**AS II** 1,548 bp	**APRT** 1,129 bp	**PRAI + InGPS** 1,932 bp		**TS α** 1,018 bp	**TS β** 1,275 bp
	yeasts (*Saccharomyces cerevisiae*)	**AS I** 1,524 bp	**AS II +**	**APRT** 1,143 bp	**PRAI** 675 bp	**+InGPS** 1,455 bp	**TS α + TS β** 2,124 bp	
	fungi (*Ustilago maydis*)	**AS I** 1,536 bp	**AS II +**	**APRT** 1,473 bp	**+ PRAI + InGPS** 2,478 bp		**TS α + TS β** 2,154 bp	
	α-proteobacteria (*Rhodopseudomonas palustris*)	**ASα + ASβ** 2,163 bp **ASα*** 1,446 bp	**ASβ*** 612 bp	**APRT** 1,017 bp	**PRAI** 657 bp	**InGPS** 798 bp	**TSα** 837 bp	**TSβ** 1,367 bp
	cyanobacteria (*Anabena variabilis*)	**ASα + ASβ** 2,187 bp **ASα** 1,515 bp	**ASβ** ‡ 555 bp	**APRT** 1,089 bp	**PRAI** 633 bp	**InGPS** 843 bp	**TS α** 801 bp	**TS β** 1,137 bp

pt Putatively targeted to plastid. Targeting to the primary plastid was predicted by TargetP. Bipatrite targeting sequences were predicted by Signal P (ER signal peptide) and TargetP (transit peptide).

M Putatively targeted to mitochondrion (TargetP prediction).

* Separated subunits of AS are encoded in some α-proteobacteria (e.g. *Erythrobacter litoralis, Silicibacter pomeroyi, Lokktanella vestfoldensis*) and are unrelated to AS fusion found in other α-proteobacteria, some cyanobacteria and diatoms.

Gene fusions are in grey, genes labeled by ▶ are fused togeteher, although they do not appear as a neighbors in the table.

‡ *A. variabilis* lack the **ASβ**. The mentioned gene coding for separate **ASβ** is from *Nostoc* sp. Gene content is very variable in cyanobacteria.

constituent genes. However, studies on bacterial trp genes and their protein products have shown that various mechanisms regulate tryptophan biosynthesis at the transcriptional, translational, and posttranslational levels (Caligiuri and Bauerle 1991). Since an extremely large amount of ATP (78 mol) is required to synthesize 1 mol of tryptophan, such strict regulation of tryptophan synthesis is advantageous to the cell, as this represents twice the energy required for the synthesis of any other amino acid (Atkinson 1977).

In contrast, the genes involved in tryptophan biosynthesis are scattered throughout the genome in all eukaryotic organisms studied so far, suggesting that each gene must be coordinately regulated by a common mechanism (Braus 1991). Over the past decades, it has mainly been the tryptophan biosynthetic pathway of bacteria and fungi that has been characterized extensively. However, many eukaryotic enzymes tend to form fusions and those of the tryptophan pathway are no exception. In yeasts (Endomycetales), component II (β subunit) of AS

is fused together with InGPS (Prantl et al. 1985), and the α and β subunits of TS are fused as well. The rest of the enzymes are encoded separately (Bartholmes et al. 1979). In other fungi, component II of AS is fused with InGPS and PRAI to form a complex called trp1 (tryptophan biosynthesis protein), and the α and β subunits of TS are also fused (DeMoss and Wegman 1965; Matchett and DeMoss 1975) (see Table 1).

Little is known about tryptophan biosynthesis in photosynthetic eukaryotes, such as land plants and algae. However, it is clear that the pathway has been relocated to the primary chloroplast (Radwanski and Last 1995). The pathway products are precursors for the synthesis of plant hormones such as indole acetic acid (IAA), phytoalexins, glucosinolates, and indole- and antharanilate-derived alkaloids. The plant enzymes are all monofunctional and all identified domains correspond to those found in microbial homologues (Radwanski and Last 1995). A single exception is represented by the large (β) subunit of AS in plants such as *Ruta graveolens* (Bohlman et al. 1995) and *Dianthus caryophylus* (Matern 1994), which produce secondary metabolites directly from anthranilate. The pathway is located in the chloroplasts of plants, with all involved enzymes being separately nuclear-encoded and posttranslationally targeted to the plastid (Zhao and Last 1995). The synthesis in rhodophytes is homologous, with the single exception of the α subunit of TS, which is still encoded in the algal chloroplast genome and not in the nucleus (Ohta et al. 1993). It is also known that TS β subunits constitute two separate clusters, TrpEb_1 and TrpEb_2. Most of the enzymes in eukaryotes belong to the TrpEb_1 cluster. Within eukaryotes, only plants are known to contain Trp_Eb_2. It has been suggested that the sole function of TrpEb_2 could be to catalyze the serine deaminase reaction (Xie et al. 2001). Nothing is known about tryptophan synthesis in photosynthetic chromalveolates possessing complex chloroplasts, such as cryptophytes, dinoflagellates, haptophytes, and diatoms (e.g., McFadden 2001). Here we present a complex phylogenetic analysis of the pathway for tryptophan biosynthesis in the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* and their heterotrophic relatives, the oomycetes *Phytophthora sojae* and *P. ramorum.* Gene arrangements and molecular phylogeny, as well as putative in silico intracellular localizations, are discussed.

Materials and Methods

All appropriate amino acid sequences of anthranilate synthase (AS), anthranilate phosphoribosyl transferase (APRT), phosphoribosyl anthranilate isomerase (PRAI), indole-3-glycerol phosphate synthase (InGPS), and both subunits of tryptophan synthase (TS) from archaea, bacteria including organelles, and eukaryotes were downloaded from GenBank. Sequences that are not available in GenBank such as those from the red alga *Cyanidioschyzon merolae* (http://www.merolae.biol.s.u-tokyo.ac.jp/; Matsuzaki et al. 2004), the green algae *Ostreococcus lucimarinus* (http://www.genome.jgi-psf.org/Ost9901_3/), *Chlamydomonas reinhardtii* (http://www.genome.jgi-psf.org/Chlre3/), the oomycetes *Phytophthora ramorum* (http://www.genome.jgi-psf.org/Phyra1_1/) and *P. sojae* (http://www.genome.jgi-psf.org/Physo1_1/) (Tyler et al. 2006), and the diatoms *Thalassiosira pseudonana* (http://www.genome.jgi-psf.org/Thaps3/; Armbrust et al. 2004) and *Phaeodactylum tricornutum* (http://www.genome.jgi-psf.org/Phatr2/) were obtained from other publicly available databases. All gene models found were compared to sequences in the NCBI protein database to identify catalytic domains (for details see Supplementary Data 1–3). Sequences were aligned using ClustalX (Thompson et al. 1997), the alignment was checked manually, and gaps and ambiguously aligned regions were excluded from further analysis. Based on the obtained multiple alignments, an appropriate model for amino acid substitution was chosen according to the AIC as implemented in PROTTEST (Abascal et al. 2005; Guindon and Gascuel 2003). Maximum likelihood (PhyML; Guindon and Gascuel 2003) and Bayesian (Mr. Bayes; Huelsenbeck and Ronquist 2001) methods were used to construct phylogenetic trees. ML trees were computed using a selected model chosen (WAG+G+I has been selected for all investigated genes) with discrete gamma distribution in four categories with all the parameters (gamma shape, proportion of invariants) estimated from the particular data set. Bayesian posterior probabilities and trees were calculated using MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001) with priors, chain number, and temperature set to default; aamodelpr was fixed to particular protein models of evolution chosen according to the AIC. Two parallel Markov chains were run for 2×10^6 generations, every 100th tree was sampled, and the first 5×10^5 generations were omitted from topology and probability reconstruction. In the case of genes encoding InGPS, where phylogenetic reconstructions led to contradictory results (for details see Results), we constructed nucleotide alignments to investigate phylogeny of this gene in more detail. We loaded InGPS nucleotide sequences into the Megalign program (DNA Star), translated nucleotide sequences into amino acids (AA), aligned AA sequences using ClustalW algorithm (as implemented in the Megalign, DNA Star), and retranslated AA back into nucleotides. An additional dataset was made by exclusion of the third codon positions. ML trees were computed from nucleotides using a model chosen by Modeltest (Posada and Crandall 1998): HKY 85 (Hasegawa et al 1985) for datasets containing all three codon positions,

and the TN93 (Tamura and Nei 1993) model for inferring phylogeny from datasets with excluded third codon positions. Both nucleotide datasets were also used to compute distance trees using LogDet/paralinear distances (Lockhart et al. 1994) as implemented in PAUP.

Homologues of the hypothetical protein COG4398, which was found to be fused with InGPS (4) in *P. tricornutum* (see Results and Discussion for details), were searched using protein BLAST (pblast) in NCBI and other databases (http://www.merolae.biol.s.u-tokyo.ac.jp/; http://www.jgi.doe.gov/). The hypothetical protein was also searched within the completely sequenced genomes using the genomic Blast search at NCBI.

All sequences of putative proteins from *P. tricornutum* were subjected to a search for N-terminal presequences known to target the enzymes into the complex diatom plastid. In the case of this pennate diatom, an EST database containing more than 120,000 ESTs is available (Maheswari et al. 2005; http://www.biologie.ens.fr/diatomics/EST/est.htm). This allows us to compare the predicted gene models to the transcript sequence, therefore making the targeting prediction more accurate. Putative ER signal peptides (SP) and transit peptides (chloroplast [cTP] and mitochondrial [mTP]) were predicted using SignalP (Nielsen et al. 1997, 1999) and TargetP (Nielsen et al. 1997; Emanuelsson et al. 2000), respectively. Since nuclear-encoded proteins are targeted to the complex plastid via the secretory pathway (Cheresh et al. 2002; Harb et al. 2004; Kroth et al. 2002), the combination of SP+TP was counted for targeting the protein to the diatom chloroplast (see Discussion for details). Although the transit peptides placed at the N terminus of most of the investigated genes were characterized using TargetP prediction as being mTP, when combined with ER SP they were evaluated as targeting the protein into the diatom chloroplast.

Results

Anthranilate Synthase

AS is an enzyme composed of two separately encoded subunits, which are both of nuclear origin in plants and algae. However, diatoms possess a very distinct fusion of both subunits, which seems to be unique in eukaryotes and unrelated to the above-mentioned plant subunits (Fig. 2, Table 1). The diatom fusion gene clusters together with fused bacterial AS, quite divergent from the other AS sequences. The phylogenetic position of the diatom AS within other bacterial homologues is unresolved due to polytomies and low support of the particular branching (see Fig. 2). However, the fusion appears only in α-proteobacteria and cyanobacteria and in a single species of *Legionella pneumophila* (γ-proteobacteria), *Syntrophomonas wolfeii* (Firmicutes), *Solibacter usitatus* (Acidobacteria), and *Thermobifida fusca* (Actinobacteria) (Fig. 2). Although there have been many genomes of Firmicutes, Actinobacteria, Acidobacteria, and γ-proteobacteria already sequenced, the AS gene fusion was found only in the above species. The topologies obtained by ML and Bayes differ slightly: the ML tree formed two monophyletic clusters (not shown), while one of the groups in the Bayesian tree appears to be paraphyletic (see Fig. 2). Notwithstanding, the basic clustering, which places the enzymes from diatoms within the bacterial AS fusions, while the sequences from oomycetes form a highly supported cluster with fungi, remained untouched (see Fig. 2 for details). It therefore appears that diatoms acquired the AS fusion by horizontal gene transfer, although the particular source of such a transfer is unclear, because of the low resolution and support of the bacterial cluster.

Anthranilate Phosphoribosyl Transferase

Based on the APRT amino acid sequences all photosynthetic eukaryotes formed a monophyletic clade placed in the sister position to the eukaryotic heterotrophs such as fungi. We can therefore propose that stramenopile APRT originates from the eukaryotic nucleus (Fig. 3). In this particular case, the diatom enzyme would originate in the nucleus of the engulfed alga (primary host), because of its highly supported clustering with oomycetes, plants, and green and red algae (see Fig. 3). This suggests that the plant-like APRT has been transferred from the red algal (primary host) nucleus to the secondary host nucleus (the diatom + oomycete ancestor) during secondary endosymbiosis. Bayesian and ML topologies were almost the same, with the exception of proteobacterial sequences forming polytomies in Bayesian trees, while in ML trees they were paraphyletic. This does not affect our hypothesis at all, because the position of plants, algae, and stramenopiles is stable regardless of the phylogenetic method used. Based on in silico prediction it appears that the *P. tricornutum* APRT contains a putative N-terminal bipartite sequence needed for targeting the protein to the diatom chloroplast.

N-(5′-Phosphoribosyl Anthranilate) Isomerase

The enzymes from fungi, red and green algae, plants, and diatoms cluster together. Since this group is unrelated to cyanobacteria or α-proteobacteria, we can suggest that all photosynthetic eukaryotes investigated in this study use the

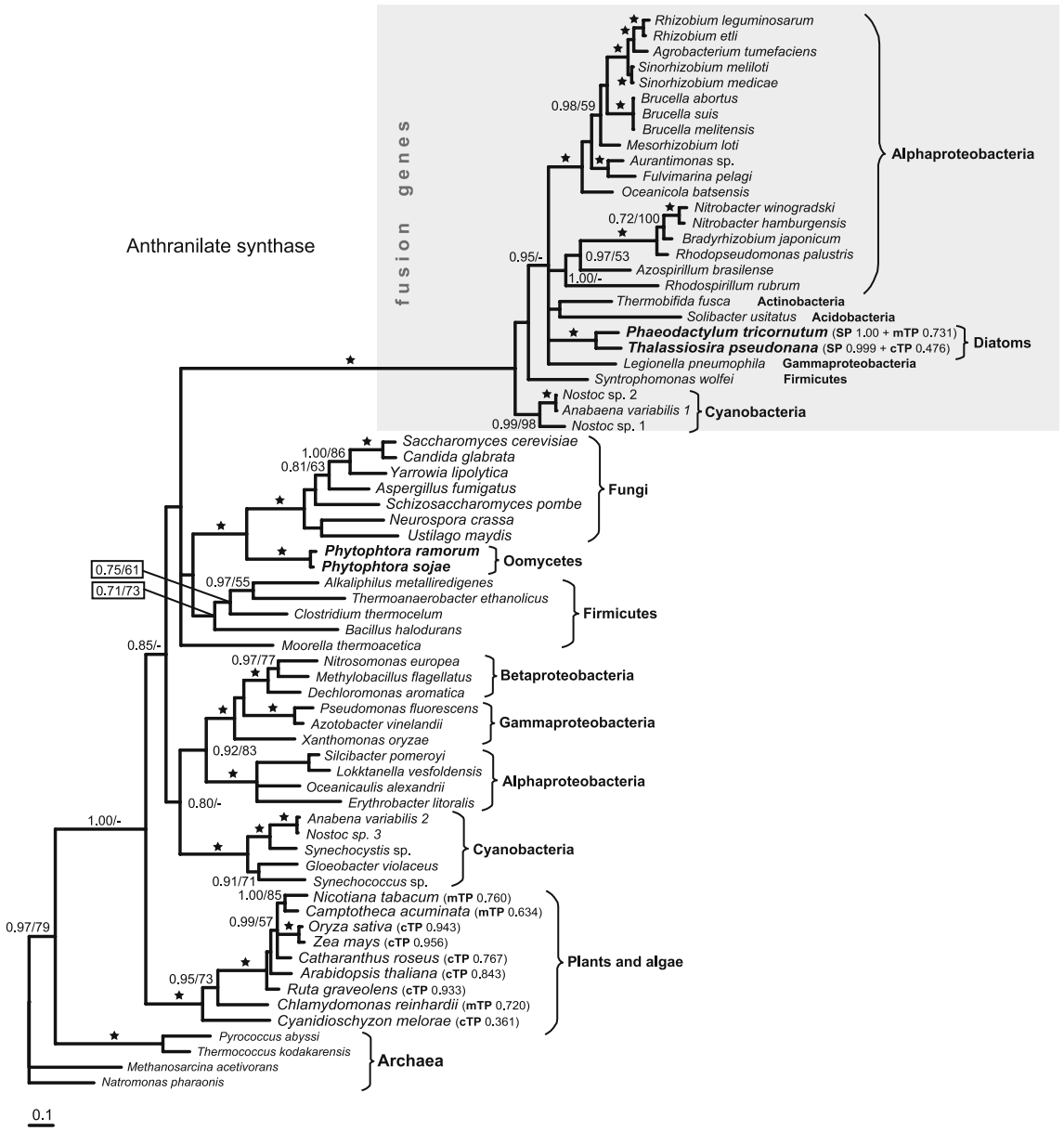


Fig. 2 Bayesian phylogenetic tree as inferred from anthranilate synthase (AS) amino acid sequences (351 positions). Numbers above branches indicate Bayesian posterior probabilities/maximum likelihood (ML) bootstrap support (WAG+I+G) in 300 replicates. Black stars indicate support by posterior probabilities of 1.00 and ML bootstraps >90%. Predicted targeting sequences are marked in the sequences of interest. SP, signal peptide; cTP, chloroplast transit peptide; mTP, mitochondrial transit peptide

PRAI that originates from the eukaryotic nucleus (Fig. 4). All stramenopiles group together with fungi with high support, forming a cluster separated from that comprising plant and algal sequences (see Fig. 4). This suggests that plants and algae have retained their original eukaryotic gene, while stramenopiles appear to use PRAI derived from the secondary host. Bayesian and ML analyses led to the same topologies. Based on in silico prediction we can suggest with high confidence that this enzyme is, at least in the case of the pennate diatom *P. tricornutum*, likely to be targeted to the chloroplast.

Indole-3-Glycerol Phosphate Synthase

Diatoms contain genes coding for InGPS of three different origins. One gene, probably of cyanobacterial origin, related to that of Bangiophyceae algae, was discovered in both diatoms (Fig. 5; *P. tricornutum* 4, *T. pseudonana* 4 The group of red algal genes, which also comprises homologues from apicomplexan parasites and is probably not related to cyanobacteria, is also present in the diatoms (Fig. 5; group E). Finally, a group containing stramenopile (secondary host) nuclear genes (Fig. 5; *P. tricornutum* 1, *T.*

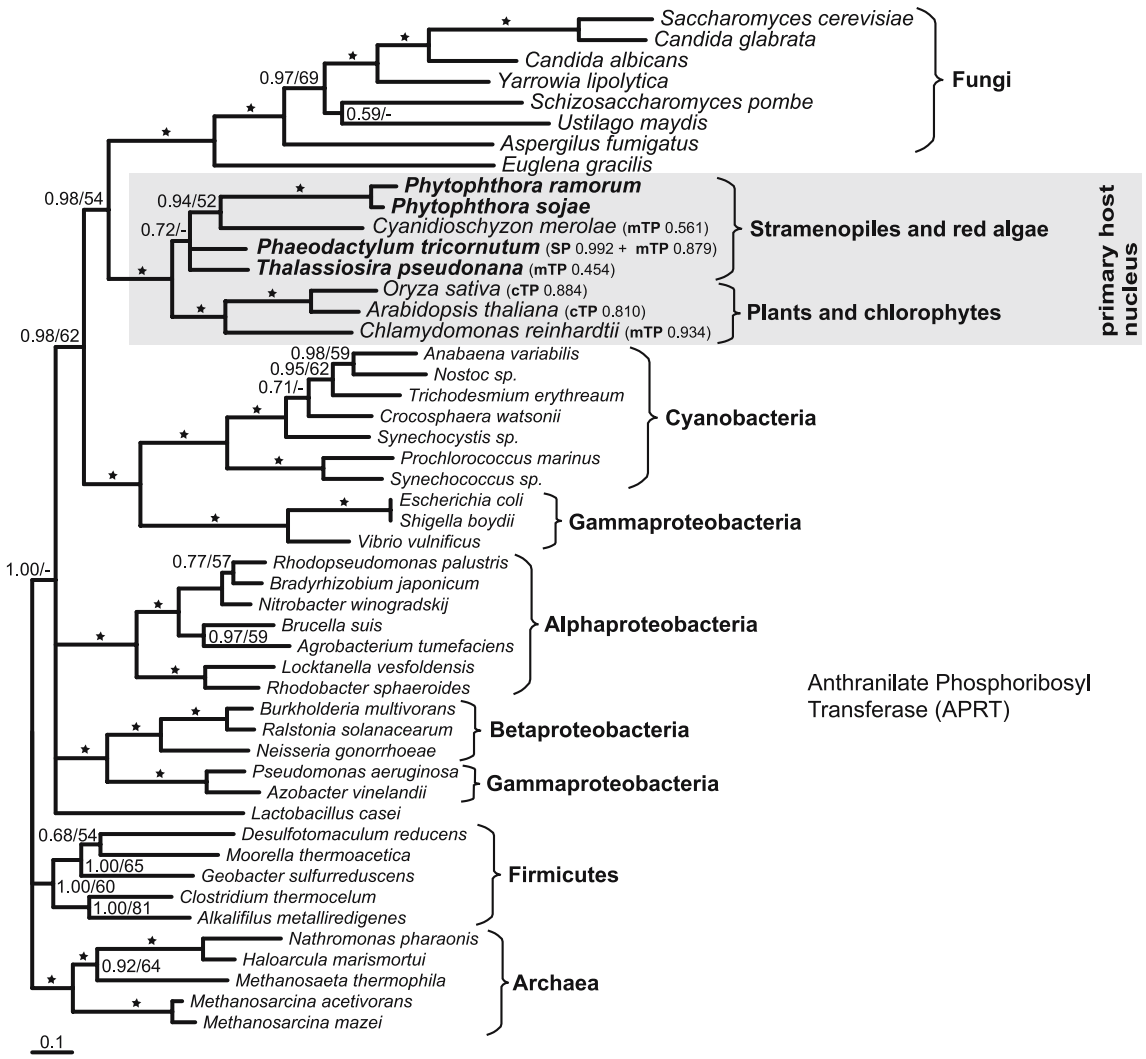


Fig. 3 Bayesian phylogenetic tree as inferred from anthranilate phosphoribosyl transferase (APRT) amino acid sequences (301 positions). The tree was computed using the WAG+I+G model for amino acid substitutions and 2 million generations. Numbers above branches indicate Bayesian posterior probabilities/maximum likelihood (ML) bootstrap support (WAG+I+G) in 300 replicates. Black stars indicate support by posterior probabilities of 1.00 and ML bootstraps >90%. Predicted targeting sequences are marked in the sequences of interest. SP, signal peptide; cTP, chloroplast transit peptide; mTP, mitochondrial transit peptide

pseudonana 1, and *P. ramorum, P. sojae*, and *P. parasitica*) that are closely related to homologues from fungi can also be detected. This clade likely represents eukaryotic genes from the secondary host (i.e., the ancestor of stramenopiles).

Although Bayesian inference displayed relationships between cyanobacterial and diatom genes coding for InGPS (Fig. 5), no other methods could confirm such a topology (ML-AA; ML computed from nucleotides and LogDet distance tree from nucleotides; see Materials and Methods for details and, also, Fig. 5B). However, the clustering of the diatom, red algal, and cyanobacterial genes was highly supported by the Bayesian posterior probability 1.00. When the other methods for tree reconstruction were used, they each gave different and extremely unsupported results (trees not shown; see Fig. 5B for details). However, it is evident that *P. tricornutum* InGPSs appeared in three different groups in the tree and formed quite long branches in two of them (Fig. 5; groups A and E). The occurrence of long branches can substantially influence the accuracy of phylogenetic analysis, because they tend to cluster together with no respect to the true relationship among taxa (Siddall and Whiting 1999). The lengths of branches indicate that all diatom sequences except those originating in the secondary host nucleus (Fig. 5; clade F) are highly divergent. Surprisingly, all of the *P. tricornutum* InGPSs, including those clustering together with fungi and oomycetes (Fig. 5; group F), contain bipartite targeting presequences at the N-terminus, suggesting their targeting

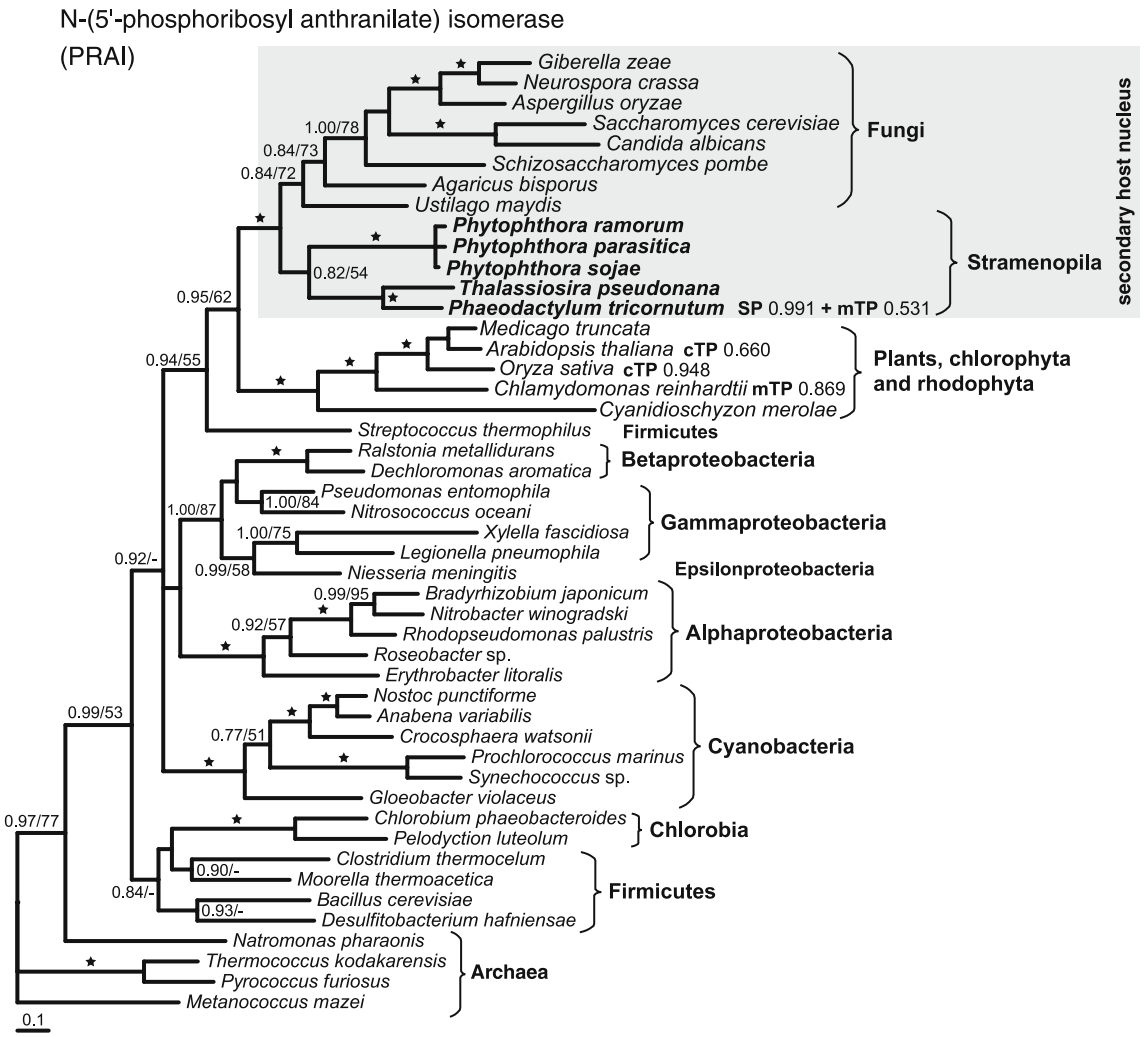
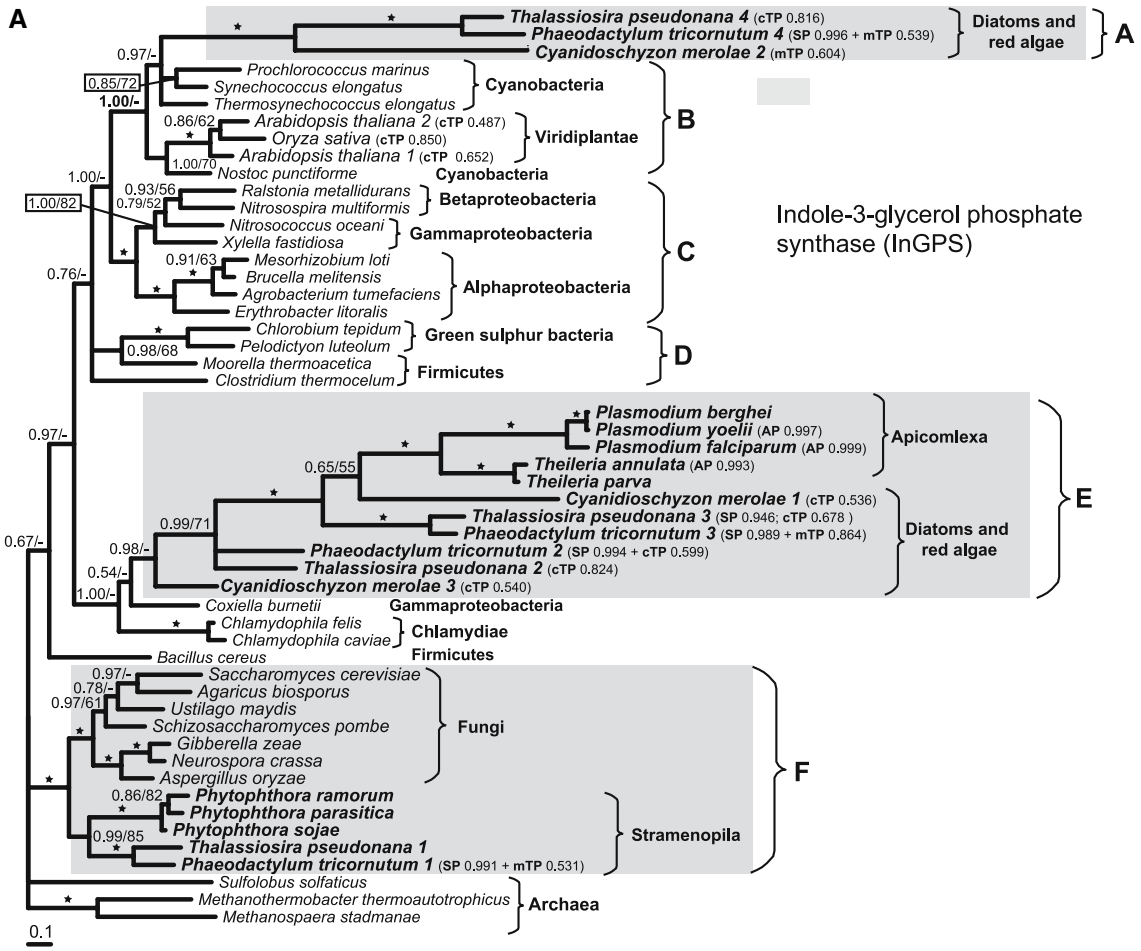


Fig. 4 Bayesian phylogenetic tree as inferred from *N*-(5′-phosphoribosyl anthranilate) isomerase (PRAI) amino acid sequences (177 positions). Numbers above branches indicate Bayesian posterior probabilities/maximum likelihood (ML) bootstrap support (WAG+I+G) in 300 replicates. Black stars indicate support by posterior probabilities of 1.00 and ML bootstraps >90%. Predicted targeting sequences are marked in the sequences of interest. SP, signal peptide; cTP, chloroplast transit peptide; mTP, mitochondrial transit peptide

to the diatom chloroplast (see Fig. 5 and Table 1). This second group of genes (Fig. 5; group E) obviously originates from the engulfed red alga. However, this group constitutes a cluster separated from the other eukaryotic genes, and displays only an unsupported affiliation to the γ-proteobacteria *Coxiella burnetii* and chlamydiae. Therefore we cannot reliably deduce whether the genes were transferred to the diatom nucleus from the red algal nucleus, the mitochondrion, or the chloroplast (Fig. 5). This clade (Fig. 5; group E) contains two genes from red algae apparently originating from gene duplication. The duplication pattern is also seen in diatoms, where each species contains two genes encoding InGPS from this separate cluster (see Fig. 5).

Since phylogenetic analyses based on InGPS are not very convincing, we have searched for another character supporting or rejecting a cyanobacterial origin for *P. tricornutum* InGPS gene 4 (Fig. 5). Interestingly, we have discovered that *P. tricornutum* InGPS is fused (no. 4; fusion is 2979 bp long; see Table 1) with a hypothetical protein (COG4398) of unknown function at the N-terminus. When examined by Blast (pblast) against the NCBI database, this hypothetical protein gave 30 hits, with the first 20 being from cyanobacteria with E values from 2e-15 to 0.009. E values of noncyanobacterial hits started at 9e-05 for a δ-proteobacterium *Anaeromyxobacter* sp. FW109-5 and at 0.033 for other noncyanobacterial sequences. The last hit *(Blastopirellula marina*) had an E value of 9.8. Surprisingly, this



B

tree	position of cluster A	support
Bayes (AA-WAG)	(((((A,B),C),D),E),F);	1.00
ML (AA-WAG)	((((B,C),D),E,A),F);	30
ML (NT-HKY)	(((((B,C),D),F),A),E);	15
ML (NT-3rd-TN93)	((((B,C),D),A,E),F);	15
ML (NT-LogDet)	((((A,D),F),(B,C)),E);	0 (?)
ML (NT-3rd-LogDet)	((((B,C),(A,D1),D2),F),E);	0 (?)

Fig. 5 Bayesian phylogenetic tree as inferred from indole-3-glycerol phosphate synthase (InGPS) amino acid sequences (228 positions). Numbers above branches indicate Bayesian posterior probabilities/ maximum likelihood (ML) bootstrap support (WAG+I+G) in 300 replicates. Black stars indicate support by posterior probabilities of 1.00 and ML bootstraps >90%. Predicted targeting sequences are marked in the sequences of interest. SP, signal peptide; cTP, chloroplast transit peptide; mTP, mitochondrial transit peptide

fusion, as well as hypothetical protein COG4398, is absent from the centric diatom *T. pseudonana*. Moreover, the COG4398 gene is absent from all photosynthetic eukaryotes we have investigated (*C. merolae, C. reinhardtii, O. lucimarinus, O. tauri, Arabidopsis thaliana, Oryza sativa, T. pseudonana*). When we searched within the completely sequenced prokaryotic and eukaryotic genomes at NCBI, we found no hits in eukaryotes (*P. tricornutum* is not covered by NCBI genomic blast). In fact homologues of COG4398 were found only in cyanobacteria.

Tryptophan Synthase

TS is a dimer composed of α and β subunits. In diatoms, both subunits are fused together and possess bipartite putative N-terminal presequences necessary for plastid targeting. However, this enzyme fusion is unique to diatoms and is not shared by any other eukaryote including the closely related oomycetes. The highly supported clustering of TSα from diatoms, oomycetes, and fungi suggests that the diatom and oomycetous enzymes originated in the

eukaryotic nucleus (see Fig. 6 for details). Because plants and algae use not the eukaryotic, but the cyanobacterial, homologue (see Fig. 6), we can suggest that TS originates in the secondary host (stramenopile) nucleus. While the centric diatoms represented here by *T. pseudonana* possess only the above-mentioned TS fusion, the pennate diatom *P. tricornutum* encodes, in addition to the fusion enzyme, an extra β subunit of tryptophan synthase. This belongs to TrpEb_1 (Xie et al. 2001); it is placed on the root of the TrpEb_1 cluster and clusters with high support together with homologues from the apicomplexan parasites *Cryptosporidium parvum* and *C. hominis* (see Fig. 7 for details). Unlike the TS fusion gene, the single β subunit does not contain any targeting sequences at the N-terminus of the protein and therefore is likely to be located in the cytosol.

Discussion

Tryptophan is an essential amino acid, which must be synthesized by an organism or obtained from the environment. Its biosynthesis is limited to photoautotrophs and a very few heterotrophs such as fungi (Fungi; Opisthokonts) and oomycetes (Stramenopiles; Chromalveolata), quite distantly related groups of organisms. Stramenopiles (Chromalveolata) contain all the enzymes of the pathway with several gene fusions including some that are absent in other eukaryotes. The pennate diatom *P. tricornutum* contains four gene fusions: one is specific to diatoms (ASα+ASβ) and some bacteria, the second is present in the nuclei of secondary host and fungi (PRAI+InGPS), and the third is found in diatoms and fungi (TSα+TSβ) but is absent from

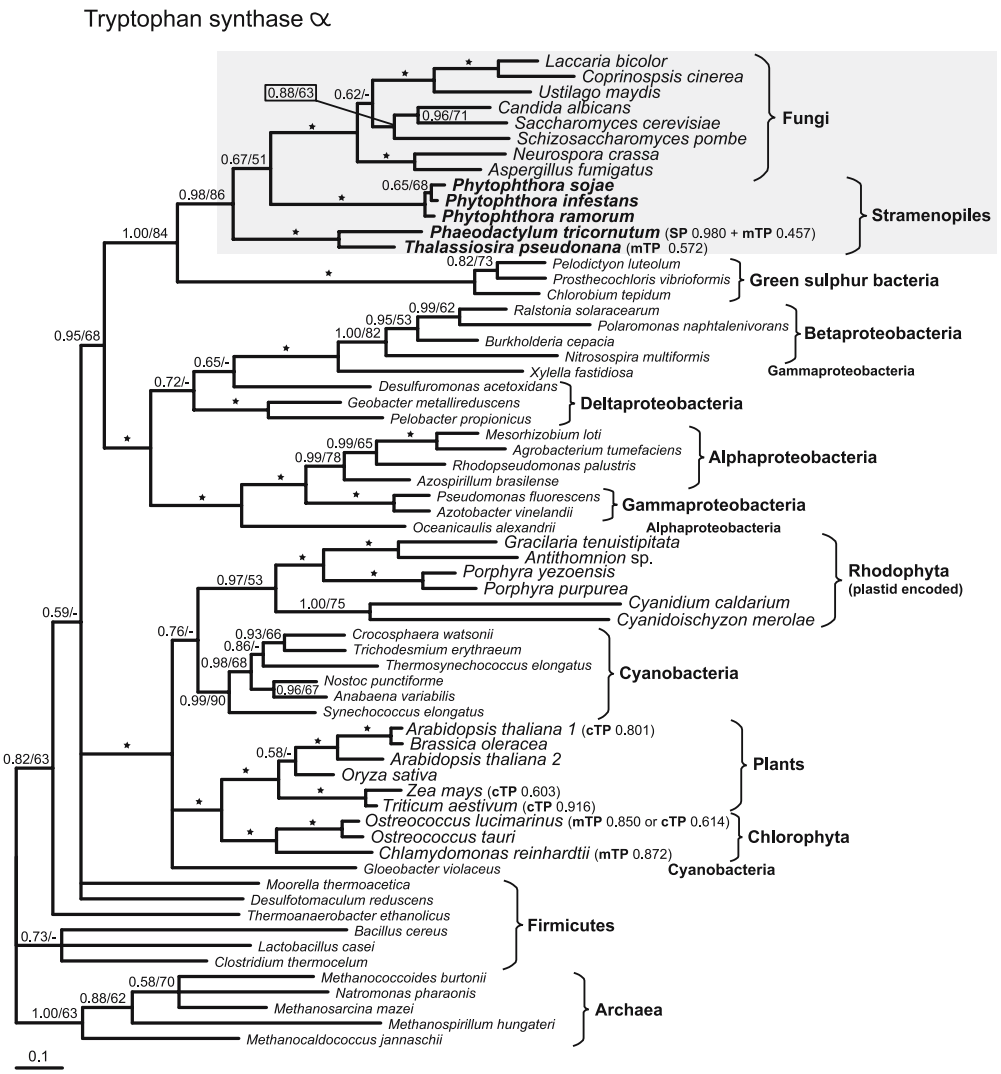


Fig. 6 Bayesian phylogenetic tree as inferred from tryptophan synthase α-subunit amino acid sequences (227 positions). Numbers above branches indicate Bayesian posterior probabilities/maximum likelihood (ML) bootstrap support (WAG+I+G) in 300 replicates. Black stars indicate support by posterior probabilities of 1.00 and ML bootstraps >90%. Predicted targeting sequences are marked in the sequences of interest. SP, signal peptide; cTP, chloroplast transit peptide; mTP, mitochondrial transit peptide

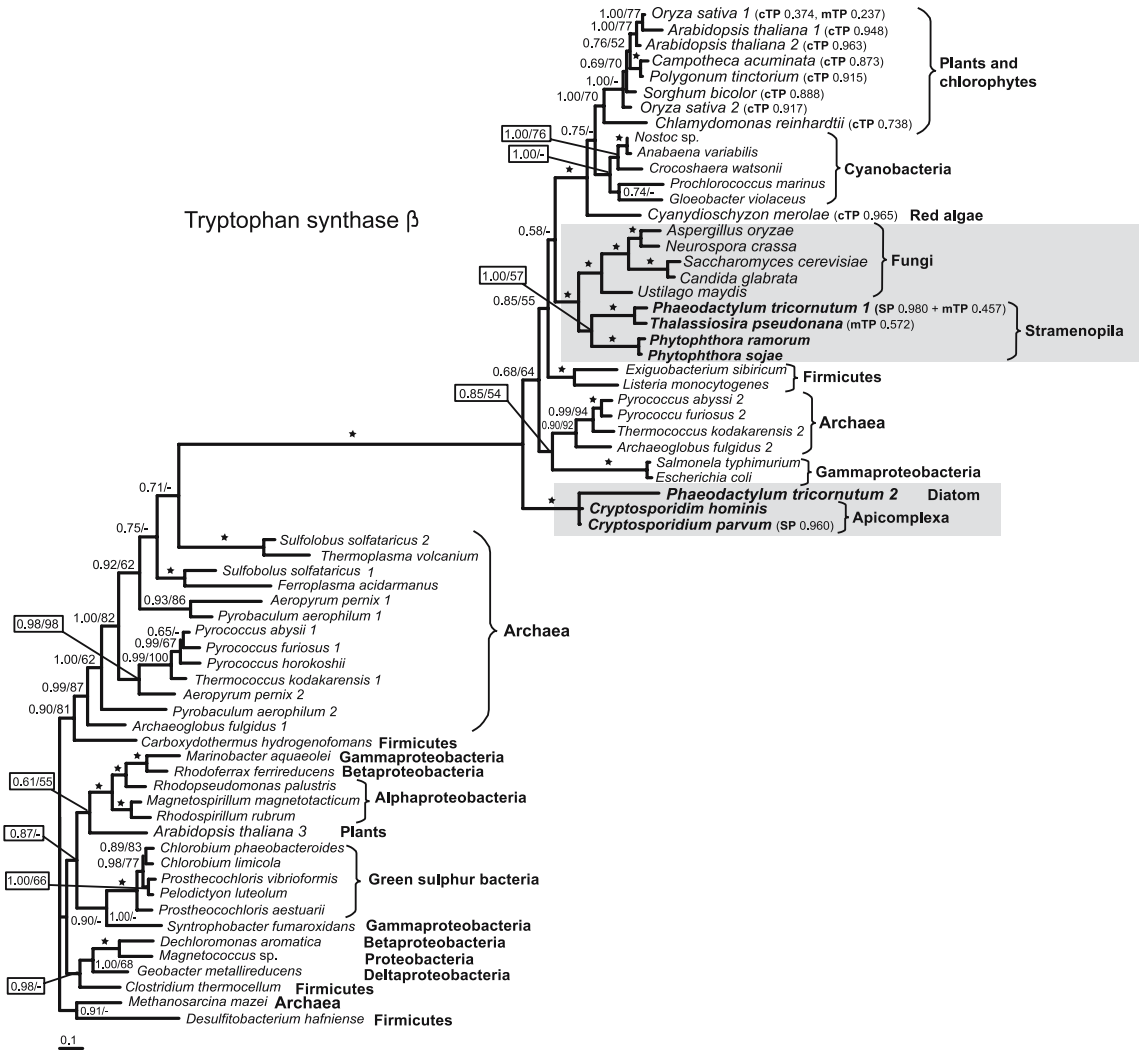


Fig. 7 Bayesian phylogenetic tree as inferred from tryptophan synthase *β*-subunit amino acid sequences (347 positions). Numbers above branches indicate Bayesian posterior probabilities/maximum likelihood (ML) bootstrap support (WAG+I+G) in 300 replicates. Black stars indicate support by posterior probabilities of 1.00 and ML bootstraps >90%. Predicted targeting sequences are marked in the sequences of interest. SP, signal peptide; cTP, chloroplast transit peptide; mTP, mitochondrial transit peptide

oomycetes (see Table 1, Fig. 8 and Fig. 9 for details). In addition to these fusions, we have discovered a fourth fusion in *P. tricornutum*, which seems to be specific for this species and is absent from all investigated photosynthetic eukaryotes (red and green algae, plants, and centric diatoms). It contains InGPS (no. 4; see Fig. 5 and Table 1) and the hypothetical protein COG4398 at the N-terminus, with the closest Blast hits found in cyanobacteria (see Results). Moreover, not even COG4398 has been found in investigated photosynthetic eukaryotes (see Results). It is important to note that when we searched for COG4398 within the completely sequenced 767 prokaryotic and 75 eukaryotic genomes, this hypothetical protein was only found in cyanobacteria, and definitely not in any eukaryotic genome. Nonetheless, even within cyanobacteria we have not found the *P. tricornutum* specific fusion, and both genes were encoded separately. Although it could appear artifactual, the presence of a regular bipartite targeting sequence at the N-terminus of the fused protein (i.e., at the N-terminus of COG4398) and the fact that the fusion is supported by ESTs almost certainly reject this possibility.

The arrangement of genes encoding the enzymes involved in the synthesis of tryptophan is very complex. Some of the analyzed proteins, such as the fusion of InGPS+PRAI found in diatoms, oomycetes, and fungi, suggest that the hypothetical ancestor of chromalveolates shared before the acquisition of the complex plastid cytosolic tryptophan biosynthesis with fungi, with the exception of AS II not being fused with InGPS+PRAI (see Fig. 9 and Table 1). However, the appearance of the ASII+InGPS fusion in yeasts (PRAI is absent, see Table 1), and the fact that fungi and diatoms share the TS fusion, unlike their

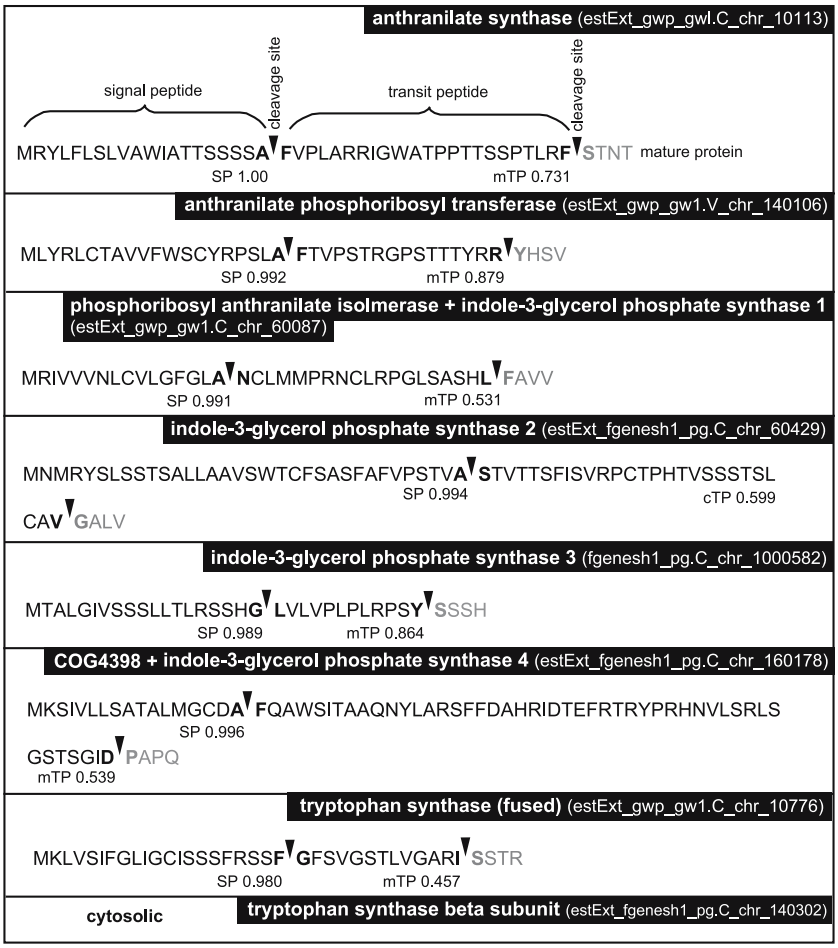


Fig. 8 Targeting presequences in tryptophan biosynthetic genes in *P. tricornutum*

close oomycete relatives, complicate the evolutionary pathway. Therefore the scenario we propose here involves fusion of AS II to InGPS+PRAI in fungi, later fission of PRAI and AS II+InGPS in yeasts, and fission of the TS subunits in oomycetes (Fig. 9). This is based on the assumption that the fungal set of genes and fusions represents an ancestral state in eukaryotes, with the exception of AS II not being fused to InGPS (Fig. 9). The supposed ancestral state is composed of separated components I and II of AS and APRT, fusion of InGPS+PRAI, and fusion of both subunits of TS (see Fig. 9). It has been inferred that gene fission is four times less probable than gene fusion (Kummerfeld and Teichmann 2005). However, it has also been suggested (Kummerfeld and Teichmann 2005) that repeated fusion or fission of the same gene is very rare in evolution. Therefore our preferred scenario has no repeated fusions, instead having two independent fusions of TSs in fungi and diatoms, and InGPS+PRAI in fungi and stramenopiles (see Fig. 9).

Genes coding for APRT represent an important molecular marker to infer the evolutionary history of stramenopiles. In our phylogenetic analysis based on APRT amino acid sequences, the genes from oomycetes, diatoms, rhodophytes, and plants constitute a common cluster, which is distinct from the cluster comprising the fungi (Fig. 3). The sister positions of these clusters, together with the lack of any relationship to proteobacteria, suggest a eukaryotic origin of the stramenopile APRT. The fact that plants, diatoms, oomycetes, and rhodophytes constitute a cluster separated from fungi demonstrates that all these genes originate from the nucleus of Archaeplastida (Plantae). We therefore postulate that the original stramenopile nuclear gene for APRT has been replaced by the red algal nuclear homologue following the endosymbiotic gene transfer (e.g., see McFadden 2001). This provides evidence for the complex plastid-containing ancestor of oomycetes, and at least in this particular aspect, it also supports the chromalveolate hypothesis (Cavalier-Smith 1999). During secondary endosymbiosis, not only were the genes encoding APRT replaced in the ancestor of oomycetes and diatoms by that from red algae, but also synthesis was relocated to the complex chloroplast. Three additional genes coding for InGPS may have been transferred from the red algal endosymbiont. However, they are only found in diatoms, and not in oomycetes, which retained the original stramenopile InGPS (see Figs. 3 and

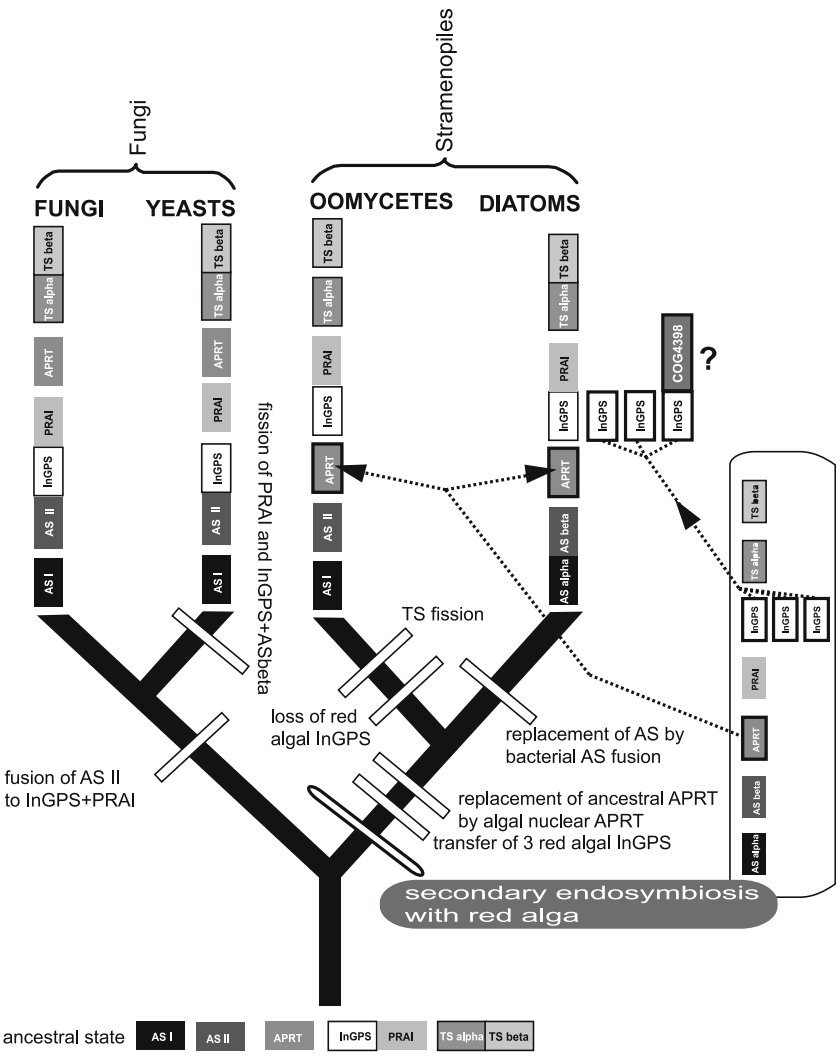


Fig. 9 Scenario for the evolution of genes involved in tryptophan biosynthesis in stramenopiles. The following arrangement is supposed to be an ancestral state: AS I, AS II, APRT, InGPS+PRAI, TSα+TSβ. In fungi, AS II fused to InGPS+PRAI; in yeasts, gene fission of PRAI occurred. In stramenopiles, after secondary endosymbiosis, three red algal InGPS genes were transferred, and ancestral APRT was replaced by that from the red algal nucleus. In oomycetes, red algal InGPSs were lost with the loss of the chloroplast, and TS fission occurred. In diatoms, AS was replaced by a bacterial AS α+β fusion

9). This suggests that with the latter loss of the chloroplast, oomycetes lost their algal InGPSs and retained the fungal-related cytosolic pathway, with the single exception represented by nuclear red algal APRT.

Three different genes coding for InGPS were found in diatoms. It must be mentioned that phylogenetic analyses based on the InGPS amino acid and nucleotide sequences gave extremely variable results, however, they always maintained these three groups of genes well defined. The first gene seems to be the only one from the pathway in diatoms which may originate from cyanobacteria. However, the sister position of this InGPS (4) to cyanobacterial, red algal, and plant genes was only supported by Bayesian inference (Fig. 5). Other methods (ML, LogDet/paralinear distances) led to unsupported topologies, and the position of InGPS (4) was the weakest one in the tree (trees not shown). On the other hand, the occurrence of the *P. tricornutum* InGPS (4) fused with the cyanobacterial gene highly favors a cyanobacterial origin of InGPS (4), as estimated by Bayesian inference. The appearance of the *P. tricornutum* specific fusion COG4398+InGPS (4) is enigmatic and difficult to explain. It is obvious that COG4398 was obtained from cyanobacteria or from the plastid genome, because this gene seems to be present only in cyanobacteria. We have noted that neither COG4398 nor the fusion was found in any other eukaryote. However, it must be clear that the complete genomic sequence is available only for a single rhodophyte species, *C. merolae* (Matsuzaki et al. 2004). It is therefore possible that, like ferrochelatase (Obornik and Green 2005) and DAHP synthase (Richards et al. 2006), the origin of enzymes located in the chloroplast is different from that of enzymes found in red algae. This may suggest that the genes of interest were lost or replaced in red algae after the secondary endosymbiotic event. Such discrepancies could also be caused by the fact that the red alga engulfed during endosymbiosis could be very different from *C. merolae*, for which we know the genome sequence. The second gene coding for InGPS is fused with PRAI and is related to the fungal "ASII(β)+InGPS+PRAI" fusion (see Fig. 5; InGPS [1]).

Based on phylogenetic analysis it originated in the eukaryotic nucleus (secondary host). This gene represents the only InGPS in the tree that does not form long branches (Fig. 5). The origin of the last InGPS with the gene duplication in diatoms and red algae, which is closely related to the enzymes from apicomplexan parasites, cannot be specified with certainty. However, due to the low or unequal support of the basal nodes occupied by chlamydias and single γ-proteobacteria, and the fact that this particular clade contains, in addition to these two bacterial sequences, only red algae and eukaryotes thought to have undergone secondary endosymbiosis (Stramenopila and Apicomplexa), we can speculate that it may represent the nuclear gene from red algae (see Fig. 5). Unexpectedly, all four enzymes are putatively targeted to the complex chloroplast in diatoms.

Diatoms use a fused bacterial AS. It has replaced the original separately encoded stramenopile genes, which has been retained in oomycetes (Fig. 2). Such a replacement is unique for eukaryotes and the gene was probably obtained via horizontal gene transfer. Due to polytomies and lack of support of the basal branches, the position of diatom AS within other bacterial fusions is unresolved (Fig. 2). Such an arrangement of both subunits of AS being fused is present in many α-proteobacteria and in a limited set of cyanobacteria. The AS fusion can also be found in a single species of γ-proteobacteria, firmicutes, actinobacteria, and acidobacteria (see Fig. 2). Since many genomes of bacteria belonging to the above-mentioned groups have been sequenced, it is quite improbable that only the named single species would retain the fused AS. Therefore we can speculate that the fused AS gene was originally present in α-proteobacteria and, due to horizontal gene transfer, was spread to other bacterial groups and to diatoms.

It has been shown that rhodophytes encode their TSα in the chloroplast genome (Ohta et al. 1993), while TSβ is nuclear encoded and posttranslationally targeted to the chloroplast (see Fig. 6 for details). In higher plants and green algae both subunits of TS have already been transferred to the nucleus. We can speculate that this may be the reason for using eukaryotic TS in stramenopiles. Since one of the TS subunits was, at the time of algal endosymbiont acquisition, encoded in the chloroplast, its transfer to the secondary host nucleus was insufficient. However, dimers such as TS should probably not be composed of subunits derived from various origins. Therefore the fused TS of nuclear origin appears to be used in diatoms and their fusion further simplifies targeting into the plastid.

It is known that nuclear-encoded proteins are targeted to the complex chloroplasts by bipartite targeting sequences (e.g., Cavalier-Smith 1999; McFadden 1999; DeRocher 2000; Kilian and Kroth 2005; Ishida 2005). Targeting sequences from *P. tricornutum* were predicted by SignalP and TargetP, respectively. In all cases, transcripts of the sequences confirmed by ESTs were used for analyses. Although both components of targeting presequences, signal peptide (SP) and transit peptide (TP), have been identified, no conserved protein motifs typical for diatom signal peptides such as ASAF or AFAP were found (Kilian and Kroth 2005). However, all SPs were predicted with high confidence (prediction not dropping below 0.980). Since datasets for prediction of chloroplast TPs are based on sequences from higher plants, we did not necessarily expect strongly supported results. However, six of seven predicted TPs were characterized as mitochondrial rather than chloroplast ones and only one of the genes encoding InGPS possessed a typical chloroplast TP. Surprisingly, the combination SP + cTP was not predicted in the cyanobacterial InGPS, as would be expected, but in one of the genes of unknown origin (*P. tricornutum* 2). To determine the putative localization of the diatom enzymes we always used the combination of ER SP + TP. There is a substantial difference in targeting of proteins to primary versus secondary chloroplasts. In plants, the type of TP is critical for selecting the target, which can be the mitochondria or the chloroplast (e.g., Mackenzie 2005). Thus the TPs of plants are under strong selection pressure to maintain the appropriate target. Proteins dually targeted to both chloroplast and mitochondrion represent the only, albeit frequent, exceptions (Mackenzie 2005). Dual targeting of nuclear-encoded proteins to the mitochondria and chloroplast in plants indicates that both TPs can do both jobs. However, in the case of complex plastids, the TP is probably not as decisive, because it is the combination SP + TP that is critical. It is possible that in diatoms the TPs are not under strong selection to discriminate between targets, and it appears that any TP in combination with a SP can target the plastid. Since all the enzymes investigated in this study possessed putative SP and TP, we suggest their putative localization within the complex diatom chloroplast.

Tryptophan biosynthesis in stramenopiles is a mosaic pathway, which is composed of enzymes of various (eukaryotic, organellar, and bacterial) origins. Such pathways have already been inferred for many metabolic routes such as polyamine biosynthesis (Illingworth et al. 2003), heme biosynthesis (Obornik and Green, 2005), and the shikimate pathway (Richards et al. 2006). However, in the current case enzymes of cyanobacterial origins have largely been replaced by eukaryotic homologues. With the exception of cyanobacterial InGPS and the bacterial fusion of AS, all other enzymes involved in tryptophan biosynthesis in diatoms appear to have originated in the

eukaryotic nucleus. However, all of the enzymes contain putative bipartite presequences at the N-terminus and are likely to be targeted to the diatom chloroplast (Fig. 7). One single stramenopile enzyme, APRT, appears to have originated from the red algal nucleus. This further suggests the possible presence of a chloroplast in the ancestor of oomycetes and its loss during evolution (Tyler et al. 2006).

Conclusions

The pathway for tryptophan biosynthesis contains two unique gene fusions in diatoms: bacterial-like ASα+ASβ, which is not found in other eukaryotes; and COG4398+InGPS (4), which is *P. tricornutum* specific. Diatoms also share other fusions with oomycetes (PRAI+InGPS [1]) and yeasts and fungi (TSα+TSβ). Only one of the involved enzymes, InGPS (4), shows a possible origin in cyanobacteria as would be expected for the nuclear-encoded protein targeted to the plastid. All the other enzymes, with the exception of bacterial ASα+ASβ, display a likely origin in the eukaryotic nucleus. Oomycetes, close relatives to diatoms, have retained APRT originating from the red algal nucleus and use this enzyme for cytosolic tryptophan synthesis. In diatoms, the entire pathway is putatively located in the complex chloroplast.

Acknowledgments This work was supported by the Grant Agency of the Academy of Sciences of the Czech Republic, Project No. IAA500220502 and Research Plan No. z60220518, and Ministry of Education of the Czech Republic, project no: 6007665801. We thank A. Lilley for critical reading of the manuscript and Beverley R. Green for helpful discussions.

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science 306:79–86

Atkinson DE (1977) Cellular energy metabolism and its regulation. Academic Press, New York

Bae YM, Crawford IP (1990) The *Rhizobium meliloti* TrpE(G) gene is regulated by attenuation, and its product Anthranilate synthase is regulated by feedback inhibition. J Bacteriol 172(6):3318–3327

Bartholmes P, Böker H, Jaenicke R (1979) Purification of tryptophan synthase from *Saccharomyces cerevisiae* and partial activity of its nicked subunits. Eur J Biochem 102:167–172

Bodyl A (2005) Do plastid-related characters support the chromalveolate hypothesis? J Phycol 41:712–719

Bohlmann J, De Luca V, Eilert U, Martin W (1995) Purification and cDNA cloning of anthranilate synthase from *Ruta graveolens*: models of expression and properties of native recombinant enzymes. Plant J 7:491–501

Braus GH (1991) Aromatic amino acid biosynthesis in the yeast *Saccharomyces cerevisiae*: a model system for regulation of a eukaryotic biosynthetic pathway. Microbiol Rev 55:349–370

Caliguiri MG, Bauerle R (1991) Subunit communication in the anthranilate synthase complex from *Salmonella typhimurinum*. Science 252:1845–1848

Cavalier-Smith T (1999) Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J Eukaryot Microbiol 46:347–366

Cavalier-Smith T (2002) Chloroplast evolution: secondary symbiogenesis and multiple losses. Curr Biol 12:R62–R64

Cheresh P, Harrison T, Fujioka H, Haldari K (2002) Targeting the malarial plastid via the parasitoforous vacuole. J Biol Chem 227:1265–1277

Crawford IP (1975) Gene rearrangements in the evolution of the tryptophan synthetic patway. Bacteriol Rev 39:87–120

Crawford IP (1989) Evolution of biosynthetic pathway: tryptophan paradigm. Annu Rev Microbiol 43:567–600

Creighton TE, Yanofsky C (1970) Chorismate to tryptophan (Escherichia coli)—anthranilate synthetase, PR transferase, PRA isomerase, InGP synthetase, tryptophan synthetase. Methods Enzymol 17:365–380

DeMoss JA, Wegman J (1965) An eznyme aggregate in tryptophan pathway of *Neurospora crassa*. Proc Natl Acad Sci USA 54:241–247

DeRocher A, Hagen CB, Froehlich JE, Feagin JE, Parsons M (2000) Analysis of targeting sequences demonstrates that trafficking to the *Toxoplasma gondii* plastid branches off the secretory system. J Cell Sci 113:3969–3977

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

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJR (2004) The evolution of modern eukaryotic phytoplankton. Science 305:354–360

Green JM, Nichols BP (1991) *p*-Aminobenzoate biosynthesis in *Escherichia coli*. J Biol Chem 266:12971–12975

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Harb OS, Chatterjee B, Fraunholz MJ, Crawford MJ, Nishi M, Roos DS (2004) Multiple functionally redundant signals mediate targeting to the apicoplast in the apicomplexan parasite *Toxoplasma gondii*. Euk Cell 3:663–674

Hasegawa M, Kishino K, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Hütter RP, Niederberger JA, DeMoss (1986) Tryptophan biosynthetic genes in eukaryotic microorganisms. Annu Rev Microbiol 40:55–77

Illingworth C, Mayer MJ, Elliott K, Hanfrey C, Walton NJ, Michael AJ (2003) The diverse bacterial origins of the Arabidopsis polyamine biosynthetic pathway. FEBS Lett 549:26–30

Ishida K (2005) Protein targeting into plastids: a key to understanding the symbiogenetic acquisition of plastids. J Plant Res 118:237–245

Kilian O, Kroth PG (2005) Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. Plant J 41:175–183

Kroth PG (2002) Protein transport into secondary plastids and the evolution of primary and secondary plastids. Int Rev Cyt Cell Biol 221:191–255

Kummerfeld SK, Teichmann SA (2005) Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet 21:25–30

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Mackenzie SA (2005) Plant organellar protein targeting: a traffic plan still under construction. Trends Cell Biol 15(10):548–554

Maheswari U, Montsant A, Goll J, Krishnasamy S, Rajyashri KR, Patell VM, Bowler C (2005) The Diatom EST Database. Nucleic Acids Res 33:D344–D347

Matchett WH, DeMoss JA (1975) The subunit structure of tryptophan synthase from *Neurospora crassa*. J Biol Chem 250:2941–2946

Matern U. (19940 Dianthus species (Carnations): In vitro culture and biosynthesis of dianthalexin and other secondary metabolites. In: Bajaj YPS (ed) Biotechnology in agriculture and forestry. Vol 28. Medicinal and aromatic plants VII. Springer-Verlag, Heidelberg, pp 170–184

Matsuzaki M, Misumi O, Shin-I T, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Nishida K, Yoshida Y, Nishimura Y, Nakao S, Kobayashi T, Momoyama Y, Higashiyama T, Minoda A, Sano M, Nomoto H, Oishi K, Hayashi H, Ohta F, Nishizaka S, Haga S, Miura S, Morishita T, Kabeya Y, Terasawa K, Suzuki Y, Ishii Y, Asakawa S, Takano H, Ohta N, Kuroiwa H, Tanaka K, Shimizu N, Sugano S, Sato N, Nozaki H, Ogasawara N, Kohara Y, Kuroiwa T (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature 428:653–657

McFadden GI (1999) Plastids and protein targeting. J Euk Microbiol 46:339–346

McFadden GI (2001) Primary and secondary endosymbiosis and the origin of plastids. J Phycol 37:951–959

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

Nielsen H, Brunak S, von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng 12:3–9

Obornik M, Green BR (2005) Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. Mol Biol Evol 22:2343–2353

Ohta N, Sato N, Kawano S, Kuroiwa T (1993) The *trpA* gene on the plastid genome of *Cyanidium caldarium* strain RK-1. Curr Genet 25:357–361

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818

Prantl F, Strasser A, Aebi M, Furter R, Niederberger P (1985) Purification and characterization of the indole-3-glycerolphosphate synthase/anthranilate synthase complex of *Saccharomyces cerevisiae*. Eur J Biochem 146:95–100

Radwanski ER, Last RL (1995) Tryptophan biosynthesis and metabolism: biochemical and molecular genetics. Plant Cell 7:921–934

Richards TA, Dacks JB, Campbell SA, Blanchard JL, Foster PG, McLeod R, Roberts CW (2006) Evolutionary origins of the eukaryotic shikimate pathway: gene fusions, horizontal gene transfer, and endosymbiotic replacements. Eukaryot Cell 5:1517–1531

Siddall ME, Whiting MF (1999) Long-branch abstractions. Cladistics 15:9–24

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 15:4876–4882

Tyler BM, Tripathy S, Zhang XM, Dehal P, Jiang RHY, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, Chapman J, Damasceno CMB, Dorrance AE, Dou DL, Dickerman AW, Dubchak IL, Garbelotto M, Gijzen M, Gordon SG, Govers F, Grunwald NJ, Huang W, Ivors KL, Jones RW, Kamoun S, Krampis K, Lamour KH, Lee MK, McDonald WH, Medina M, Meijer HJG, Nordberg EK, Maclean DJ, Ospina-Giraldo MD, Morris PF, Phuntumart V, Putnam NH, Rash S, Rose JKC, Sakihama Y, Salamov AA, Savidor A, Scheuring CF, Smith BM, Sobral BWS, Terry A, Torto-Alalibo TA, Win J, Xu ZY, Zhang HB, Grigoriev IV, Rokhsar DS, Boore JL (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313:1261–1266

Xie G, Forst C, Bonner C, Jensen RA (2002) Significance of two distinct types of tryptophan synthase beta chain in Bacteria, Archaea and higher plants. Genome Biol 3:0004.1–0004.13

Yanofsky C, Platt T, Crawford IP, Nichols BP, Christie GE, Horowitz H, Vancleemput M, Wu AM (1981) The complete nucleotide sequence of the tryptophan operon of *Escherichia coli*. Nucleic Acids Res 9:6647–6668

Yoon HS, Hacket JD, Pinto G, Bhattacharya D (2004) Molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21:809–818

Zhao J, Last RL (1995) Immunological characterization and chloroplast import of the tryptophan biosynthetic enzymes of the flowering plant *Arabidopsis thaliana*. J Biol Chem 270:6081–6087