Eukaryotic Pyruvate Formate Lyase and Its Activating Enzyme Were Acquired Laterally from a Firmicute

Courtney W. Stairs, Andrew J. Roger, and Vladimir Hampl*

1Centre for Comparative Genomics and Evolutionary Bioinformatics, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada
2Department of Parasitology, Faculty of Science, Charles University in Prague, Prague, Czech Republic

*Corresponding author: E-mail: vlada@natur.cuni.cz
Associate editor: Charles Delwiche

Abstract

Most of the major groups of eukaryotes have microbial representatives that thrive in low oxygen conditions. Those that have been studied in detail generate ATP via pathways involving anaerobically functioning enzymes of pyruvate catabolism that are typically absent in aerobic eukaryotes and whose origins remain controversial. These enzymes include pyruvateferredoxin oxidoreductase, pyruvate:NADP⁺ oxidoreductase, and pyruvate formate lyase (Pfl). Pfl catalyzes the nonoxidative generation of formate and acetyl-Coenzyme A (CoA) from pyruvate and CoA and is activated by Pfl activating enzyme (Pfla). Within eukaryotes, this extremely oxygen-sensitive pathway was first described in the hydrogenosomes of anaerobic chytrid fungi and has more recently been characterized in the mitochondria and chloroplasts of the chlorophyte alga Chlamydomonas reinhardtii. To clarify the origins of this pathway, we have comprehensively searched for homologs of Pfl and Pfla in publicly available large-scale eukaryotic genomic and cDNA sequencing data, including our own from the anaerobic amoebozoan Mastigamoeba balamuthi. Surprisingly, we find that these enzymes are widely distributed and are present in diverse facultative or obligate anaerobic eukaryotic representatives of the archaeplastidan, metazoan, amoebozoan, and haptophyte lineages. Using maximum likelihood and Bayesian phylogenetic methods, we show that the eukaryotic Pfl and Pfla sequences each form monophyletic groups that are most closely related to homologs in firmicute gram-positive bacteria. Topology tests exclude both α-proteobacterial and cyanobacterial affinities for these genes suggesting that neither originated from the endosymbiotic ancestors of mitochondria or chloroplasts. Furthermore, the topologies of the eukaryote portion of the Pfl and Pfla trees significantly differ from well-accepted eukaryote relationships. Collectively, these results indicate that the Pfl pathway was first acquired by lateral gene transfer into a eukaryotic lineage most probably from a firmicute bacterial lineage and that it has since been spread across diverse eukaryotic groups by more recent eukaryote-to-eukaryote transfer events.

Key words: pyruvate formate lyase, lateral gene transfer, eukaryotic phylogeny, origin.

Introduction

In aerobic eukaryotes, carbohydrates are catabolized glycolytically to produce pyruvate that is converted to acetyl-Coenzyme A (CoA) in mitochondria by the pyruvate dehydrogenase complex (Pdh). Acetyl-CoA feeds into the tricarboxylic acid cycle generating reducing equivalents that ultimately drive ATP synthesis via the coupling of the respiratory chain and oxidative phosphorylation. Multicellular organisms that live in the absence of oxygen, such as anaerobic helminths, have evolved alternatives to aerobic energy production pathways such as the malate dismutation pathway (Tielens and Van Hellemond 1998) and often utilize alternative terminal electron acceptors of the respiratory chain (Kobayashi et al. 1996). Anaerobic unicellular eukaryotes (protists) that have been studied in detail also utilize pyruvate to produce ATP, often by substrate-level phosphorylation. Furthermore, they differ from aerobic eukaryotes in the enzymes they use to convert pyruvate to acetyl-CoA. Instead of using Pdh, these organisms utilize enzymes, such as pyruvate-ferredoxin oxidoreductase (Pfo), pyruvate:NADP⁺ oxidoreductase (Pno), and/or pyruvate formate lyase (Pfl).

Although Pdh is known to function exclusively within mitochondria of aerobic eukaryotes, these anaerobic enzymes are localized to different subcellular compartments depending on the organism. For example, in some organisms, Pfo functions within mitochondrion-related organelles (MROs) such as the hydrogenosomes of the parabasalid parasite Trichomonas vaginalis (Steinbuchel and Müller 1986; Hrdy and Muller 1995) and in others, such as Entamoeba histolytica and Giardia intestinalis, it is thought to primarily function within the cytosol (Embley and Martin 2006). The enzyme Pno has been shown to localize in the mitochondria of Euglena gracilis and Plasmodium falciparum and in the mitosomes of Cryptosporidium parvum (Rotte et al. 2001; Mogi and Kita 2010).

Pfl (EC 2.3.1.54), is a member of the pyruvate formate lyase protein family and catalyzes the nonoxidative generation of formate and acetyl-CoA from pyruvate and CoA. Pdh, Pfo, and Pno (and not Pfl) primarily use redox chemistry with cofactors (NAD⁺ or ferredoxins) to convert pyruvate into acetyl-CoA. Pfl activity also involves radical chemistry. However, radical generation occurs independent of catalysis and involves a second enzyme pyruvate formate lyase activating enzyme (Pfla; EC 1.14.19.3).
Pfl is activated by the formation of an α-carbon centered radical in its C-terminal glycylic radical domain (pfam01228) by Pfla which utilizes the cofactors S-adenosyl methionine (SAM) and reduced flavodoxin or ferredoxin (Wagner et al. 1992). When pyruvate binds to Pfl, the radical is transferred to two adjacent active site cysteine residues (Frey et al. 1994). These conserved cysteine residues along with the conserved tyrosine downstream relative to the glycine within the glycylic radical domain, can be used to distinguish pyruvate-catalyzing Pfls from other Pfl protein family members such as the benzyl succinate synthases and glycerol dehydratases that have only one cysteine in the active site and a different radical domain architecture (Lehtio and Goldman 2004).

Upon oxygen exposure, the radical form of Pfl upon oxygen exposure is irreversibly cleaved at the radical residue (Wagner et al. 1992). Presumably for this reason, the enzyme has only been identified in organisms that are either obligatory or transient anaerobes, such as the firmicutes (Thauer et al. 1972) and Enterobacteriaceae. Pfl activity in eukaryotes has been described in the cell-free extracts (Marvin-Sikkema et al. 1993) and hydrogenosomes of the chytrid fungus Necallimastix sp. (Akhanova et al. 1999) as well as its close relative Piromyces sp. E2 (Boxma et al. 2004), the chloroplasts and mitochondria of the chlorophyte algae Chlamydomonas reinhardtii and probably its close relative Polytomella sp. (Atteia et al. 2006; Hemschemeier et al. 2008). It has been suggested that Pfl is part of the ATP-producing pathway in these organelles. Unlike other pyruvate-metabolizing enzymes, the action of Pfl does not involve CO₂ production. Indeed, this system has been hypothesized to function as an alternative to PfO, when the latter is inhibited by high CO₂ levels in the organism’s environment (Boxma et al. 2004). However, Pfl-mediated acetyl-CoA generation is problematic for energy production because of the SAM consumed in the initial activation of Pfl; SAM requires ATP for its synthesis by SAM synthetase. Unless the energy required for SAM synthesis is regenerated elsewhere and the radical is protected, the Pfl system cannot be as efficient as a PfO-based system for ATP production. Nevertheless, it is possible that Pfl-generated acetyl-CoA is also important for sustaining biosynthetic pathways, such as fatty acid biosynthesis and elongation and amino acid synthesis.

The evolutionary origins of anaerobic pyruvate oxidizing enzymes within eukaryotes are contentious. It has been suggested that they may have originated with the α-proteobacterial mitochondriald sibomt that may have been a facultative anaerobe (Martin and Müller 1998). To evaluate this, as well as alternative hypotheses, we recently surveyed the presence and absence of PfO- and PfO-related proteins, as well as [FeFe] hydrogenases in partial and complete genomic and transcriptomic sequence data of diverse eukaryotes (Hug et al. 2010). Although the homologs of PfO and PfO proteins that were identified in diverse eukaryotic lineages appeared to form a monophyletic group, phylogenetic analyses did not support an α-proteobacterial origin for the eukaryotic enzymes. Similarly, a previous network analysis demonstrated that the Pfl sequences from Chlamydomonas and the chytrid fungi form a grouping clustering near homologs from firmicutes (Gelius-Dietrich and Henze 2004). With recent largescale genomic and cDNA/expressed sequence tag (EST) sequencing projects of eukaryotes, we have identified many additional sequences homologous to Pfl and PfO in a wide diversity of other eukaryotes including members of the Amoebozoa, Metazoa, Haptophyta, and Archaeplastida. In this study, we investigate the origin of eukaryotic Pfl and more generally the distribution and evolution of pyruvate metabolizing enzymes within eukaryotes to trace their prokaryotic ancestry.

### Methods

#### 454 EST Project of Mastigamoeba balamuthi

*Mastigamoeba balamuthi* cultures were maintained from an established American Type Culture Collection (ATCC) culture (ATCC #30984) in PYGC media (1% peptone, 1% yeast extract, 56 mM glucose, 86 mM NaCl 10 mM K₂H₂PO₄, pH 6.8) grown at 20-24 °C. Total RNA was isolated using Trizol (Tri-reagent) following the protocol supplied by the manufacturer (Invitrogen). A cDNA library was constructed by Vertis Biotechnologies AG (Freising, Germany), and 454 pyrosequencing was performed by Genome Québec. The Newbler assembly program (Roche) was used to assemble the reads. PfO and PfO were identified using basic local alignment search tool (BLAST) and candidate reads and contiguous sequences were viewed and manually assembled and edited using Sequencer (Gene Codes corp.). Sequences were confirmed by polymerase chain reaction on cDNA and genomic DNA followed by Sanger sequencing. The sequences of *M. balamuthi* PfO and PfO genes are deposited in GenBank with the accession numbers HM590578 and HQ003218, respectively.

#### Database Searches

Bacterial PfO and PfO protein sequences were retrieved from all whole-genome sequencing projects compiled in the microbial online analysis database (http://ratite.cs.dal.ca/moa/) using blastp and tblastn with PfO and PfO from *Chla. reinhardtii* (gi:92084842, 57021069) as queries. Because only pyruvate catalyzing PfO proteins were sought, those sequences lacking the diagnostic conserved active site motif (CC, Lehtio and Goldman 2004), glycylic radical domain ([LIVM]R [LIVM]SGY, Lehtio and Goldman 2004), or those annotated as PfO4 were removed. Similarly, those PfO sequences characterized as non-pyruvate PfO activating enzymes that had an additional ferredoxin domain (Raynaud et al. 2003) or from taxa without a pyruvate-catalyzing PfO were removed. To maintain a reasonable-sized data set for subsequent phylogenetic analyses, extremely similar sequences from “redundant” species and “strains” were randomly removed (see supplementary table S1, Supplementary Material online).

Eukaryotic gene and protein sequences were retrieved from a variety of sources including EST and genome databases from: the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov), the Joint
Streptococcus suis sequences (smallest percent identity observed between two Pfla sequences was identified as 26.804% and 30.928%, respectively. The parameter values calculated using RAxML (see Susko 2010).

The same procedures were performed to probe the presence/absence of Pdh, Pfo, Pno, and sulfite reductase (SR) protein families in diverse eukaryotic genomic and transcriptomic data. A Pdh complex was designated as present if at least Pdh (E1 subunit) was present in the data. Those proteins that resembled Pfo but had a fused C-terminal CysI domain (Crane et al. 1995), or cytochrome P450 reductase domain (Zeghouf et al. 1998) were classified as SR or Pno, respectively.

Data Set Generation and Phylogenetic Analysis
Amino acid sequences were aligned using MUSCLE v3.2 (Edgar 2004) and manually edited to mask out regions of ambiguous alignment. The final alignments for Pfl and Pfla consisted of 693 sites (166 taxa) and 194 sites (152 taxa), respectively. Alignments are available upon request from the authors. The lowest percent identity observed between two Pfla sequences (Ostreococcus tauri and Bifidobacterium adolescentis) before and after masking was identified as 26.804% and 30.928%, respectively. The smallest percent identity observed between two Pfl sequences (Streptococcus suis and Pirormyces sp.) before and after masking was identified as 38.318% and 39.768%, respectively. In all analyses, the maximum percent identity was greater than 99%. Maximum likelihood (ML) phylogenetic analysis was performed using RAxML version 7.2.6 (Stamatakis 2006) under the Le and Gascuel (LG) (Le et al. 1995), or cytochrome P450 reductase domain (Zeghouf et al. 1998) were classified as SR or Pno, respectively.

The program requires nonzero branch lengths and sufficient alignment data for estimation. For this reason, taxa with extremely short branch lengths were removed from the trees and alignments and simulated data sets (10,000 sites using SeqGen http://tree.bio.ed.ac.uk/software/seagen/) were used. aBPn values can be considered as a statistical test that the split was not present in the true phylogeny with a P value = 1 – BPc. Bayesian inference (BI) was also conducted using PhyloBayes 3.2 (Lartillot et al. 2009) by running two chains under the catfix C20 model of evolution (Le et al. 2008). For each chain, a total of 300,000 generations were run, from which trees were sampled every 100 generations and discarding a manually determined burn-in of 50,000 generations for each (yielding a total of 2,500 trees). Posterior probabilities (PP) for splits were mapped on to the ML estimated topology.

Congruence Testing and Concatenation
Organisms that had multiple paralogous copies of Pfl and only one Pfla were removed prior to concatenation because each Pfl could not be uniquely assigned an activating enzyme. Congruence between Pfl and Pfla phylogenies was tested for the data set without paralogues using CONCATENATOR with an alpha-level cutoff of 0.05 (Leigh et al. 2008). For concatenated Pfl–Pfla analyses, to assess potential artifacts introduced by gene and lineage-specific evolutionary dynamics, ML analysis was performed both with and without partitioning the data to allow for variable edge lengths and other parameters for each protein family (the RAxML—M option). The alpha shape parameter estimated for the concatenated unpartitioned data set was 0.697047. In general, only minor differences in bootstrap values on splits were noted between partitioned versus nonpartitioned analyses therefore the bootstrap correction and BI PP were calculated only for the nonpartitioned data set. To evaluate the impact of missing data on topological support values, a second analysis was performed as described above where only complete sequences and partial sequence with greater than 50% of the sites were included.

Topology Testing
To test alternative candidate phylogenetic positions of eukaryotes within the bacterial subtree, we used the approximately unbiased (AU) topology test implemented in CONSEL (Shimodaira and Hasegawa 2001). In brief, ML trees were estimated subject to topological constraints corresponding to alternative hypotheses (e.g., x-proteobacteria as sister to the eukaryotes) using RAxML (—g option). The AU tests were performed with CONSEL (table 1) based on these trees as well as a sample of 100 “ML-bootstrap” trees generated during bootstrapping using RAxML (—f g option). The 100 additional trees were to provide a sample of “good” trees, which is necessary in order for the P values for the AU test to be accurate (Shi et al. 2005).

Gene Order and Operon Prediction
Putative operons in bacterial genomes were identified using OperonDB (Pertea et al. 2009). If Pfl and Pfla were located at distantly separated loci on a given prokaryotic
chromosome, they were not considered to be in an operon. Complete genomes of the eukaryotes *Thalassiosira pseudonana*, *Ostreococcus* species, and *Chlorella reinhardtii* (gi) were examined for the location and order of the Pfl and Pfla genes on their assembled scaffolds.

Tests for Long-Branch Attraction, Amino Acid Composition Bias

To evaluate whether long-branch attraction contributed to the support for the grouping of the eukaryotic sequences, for each of the analyses, we: 1) removed 50% of the longest eukaryotic branches and the resulting data set was reanalyzed as described above 2) removed the fastest evolving sites, which are often responsible for long-branch attraction, and reanalyzed the truncated alignments as described above. The fast evolving sites were identified and removed using AIR-identifier and AIR-remover (Kumar et al. 2009). To evaluate the possibility of phylogenetic artifacts associated with amino acid composition bias within the eukaryotic sequences, we performed a chi-squared test for deviations in amino acid compositions for each taxon implemented in TREE-PUZZLE version 5.2 (Schmidt et al. 2002). Taxa with significantly different amino acid composition from the overall data set frequencies were removed, and the analysis was repeated as above to check whether estimated topologies and support values changed. Additionally, we performed cluster analyses of the amino acid composition vectors of the eukaryotic sequences and the concatenated analysis. Euclidean distance matrices between composition vectors were constructed from 100 bootstrapped alignments and used to generate 100 UPGMA trees. The majority rule consensus of these trees and bootstrap support for bipartitions were visualized using CONSENSE (Felsenstein 2005).

<table>
<thead>
<tr>
<th>Phylogeny</th>
<th>Monophyly Test</th>
<th>AU Test P value</th>
<th>AU Test P value of Partitioned Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfl</td>
<td>Eukaryotes and γ-proteobacteria</td>
<td>0.0004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eukaryotes and α-proteobacteria</td>
<td>0.0024</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eukaryotes and α-clade</td>
<td>0.0130</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eukaryotes and cyanobacteria</td>
<td>0.1580</td>
<td></td>
</tr>
<tr>
<td>Pfla</td>
<td>Eukaryotes and γ-proteobacteria</td>
<td>0.8010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eukaryotes and α-proteobacteria</td>
<td>0.0010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eukaryotes and α-clade</td>
<td>0.4660</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eukaryotes and cyanobacteria</td>
<td>0.0740</td>
<td></td>
</tr>
<tr>
<td>Concatenated tree (Pfl–Pfla)</td>
<td>Eukaryotes and α-proteobacteria</td>
<td>2 × 10⁻³²</td>
<td>5 × 10⁻⁹⁹</td>
</tr>
<tr>
<td></td>
<td>Eukaryotes and α-clade</td>
<td>0.0130</td>
<td>0.0040</td>
</tr>
<tr>
<td></td>
<td>Eukaryotes and cyanobacteria</td>
<td>1 × 10⁻⁵</td>
<td>0.0090</td>
</tr>
<tr>
<td>Concatenated tree (Pfl–Pfla; exclusively eukaryotes)</td>
<td>Typical eukaryotic phylogeny</td>
<td>0.0020</td>
<td>0.0030</td>
</tr>
<tr>
<td></td>
<td>Chlorophytes and prasinophytes</td>
<td>0.0100</td>
<td>0.0080</td>
</tr>
<tr>
<td></td>
<td>Chlorophytes, prasinophytes, and <em>Mastigamoeba</em></td>
<td>0.0380</td>
<td>0.0330</td>
</tr>
<tr>
<td></td>
<td>Fungi and prasinophytes</td>
<td>0.5230</td>
<td>0.4660</td>
</tr>
<tr>
<td></td>
<td>Opisthokonts</td>
<td>0.6610</td>
<td>0.6650</td>
</tr>
</tbody>
</table>

* AU topology tests for the single gene and concatenated Pfl and Pfla data sets.

In Silico Subcellular Localization Prediction

TargetP (available at http://www.cbs.dtu.dk/services/TargetP/ (Emanuelsson et al. 2000) and Mitoprot II (available at http://ihg2.helmholtz-muenchen.de/ihg/mitoprot.html (Claros and Vincens 1996)) were used to predict subcellular localization of proteins based on their amino acid sequences.

Results and Discussion

Identification of Previously Unidentified Pfl and Pfla in Eukaryotes

Database mining of all publicly available genome and EST projects for eukaryotes revealed a large number of previously unreported Pfl and Pfla homologues (fig. 1). A total of 16 new Pfl eukaryotic homologues were identified in archaeoplastids (*Scenedesmus obliquus*, *Chlorella* sp., *Acetabularia acutabulum*, *Haematococcus pluvialis*, *Volvox carteri*, *Os. tauri*, *Os. lucimarinus*, *Micromonas pusilla*, *Micromonas sp.*, *Porphyra haitanensis*, and *Cyanophora paradoxa*), an opisthokont (*Amoebidium parvum*), an amoebozoan (*Ma. balamuthi*), a stramenopile (*Tha. pseudonana* (2)), and a haptophyte (*Prymnesium parvum*). A total of seven new Pfla eukaryotic homologues were identified in archaeoplastids (*Mi. pusilla*, *Micromonas sp.* Os. *tauri*, and Os. *lucimarinus*), an amoebozoan (*Ma. balamuthi*), and a stramenopile (*Tha. pseudonana* (2)). The majority of these sequences were identified in both genomic and EST data. Furthermore, in most cases, spliceosomal introns were identified in the genomic sequences confirming a eukaryotic (rather than bacterial) provenance. Pfl, but not Pfla, sequences were also found in the meta-zoa *Brugia malayi* (a filarial nematode) and *Litopenaeus vannamei* (a whiteleg shrimp) genomic databases. The
**Brugia** genomic sequence corresponding to Pfl was identified in an incomplete genome project, did not contain spliceosomal introns and was not identifiable in a transcriptome project (http://www.nematodes.org). The **Lit. vannamei** sequence was extremely short (~500 bp), only identified in EST data and 74% identical at the nucleotide level to **Listeria innocua** pfl. Although it remains uncertain, for these reasons, we suspect both sequences are bacterial contaminants and were not included into subsequent analyses.

**Eukaryotic Monophyly of Pfl and Pfla**

The global phylogenetic analyses of all representative bacterial and eukaryotic pyruvate-catalyzing Pfl and Pfla amino acid sequences are shown in supplementary figures S1 and S2, Supplementary Material online, respectively. In the Pfl
analysis, all eukaryote sequences form a clade with moderate support (BP = 73, BPc = 88, PP = 0.66), showing a weak affinity (<50% BP) for a firmicute plus Bacteroides grouping. Similar results were recovered for the Pf a phylogeny where eukaryotes were monophyletic with significant support (BP = 90, BPc = 97, PP = 1.0) (see supplementary fig. S2, Supplementary Material online). To test whether eukaryotic monophyly in these phylogenies were artifacts of long-branch attraction, the analyses were repeated with 50% of the longest eukaryotic branches removed (data not shown). This secondary analysis did not change the overall tree topology, however it did differentially affect the support for eukaryote monophyly in the Pf and Pf a analyses (BPf = 79, PPf = 1.0; BPf = 83, PPf = 1.0). For the Pf and Pf a analyses, there is very weak support for intergroup relationships especially along the “backbone” of the phylogeny (i.e., the innermost branches on the tree). Although we initially suspected this could be due to the position of the strongly supported γ-proteobacterial clade relative to other sequences (this is the deepest significant split in the tree), removal of the γ-proteobacteria did not alter the resolution amongst the other taxa (data not shown).

The most common bacteria-related sequences in eukaryotic genomes are the genes that were acquired during the establishment of endosymbiologically-derived mitochondria and chloroplasts (Pisani et al. 2007). To test whether the position of the monophyletic eukaryotic group in the Pf and Pf a trees was consistent with acquisition during either of these endosymbioses, we tested whether ML topologies obtained by constraining the eukaryotic sequences to group with α-proteobacteria (mitochondria) or cyanobacteria (chloroplasts) respectively were significantly rejected by the data (table 1). Eukaryotes plus α-proteobacteria monophyly was rejected for both Pf (AU P value = 0.002) and Pf a (P = 0.001). As we noted that the α-proteobacteria often grouped with the β-proteobacteria and actinobacteria in optimal trees (α-clade, supplementary fig. S1, Supplementary Material online), we tested whether the eukaryotic affinity for this group as a whole; this hypothesis was rejected by Pf (P = 0.013) but not by Pf a (P = 0.466) data sets (table 1). The chloroplast origin hypothesis that constrained the cyanobacteria to group with the eukaryotes was not rejected for either Pf (P = 0.158) nor for Pf a (P = 0.074).

Although a cyanobacterial/chloroplast origin cannot be rejected in these analyses, it seems rather unlikely because a number of the Pf and Pf a-containing eukaryotes, such as Ma. balamuthi, the chytrid fungi, and Am. parasiticum are members of the so-called “unikont” eukaryotes that are likely to have diverged from eukaryotes prior to the origin of chloroplasts within eukaryotes (Lane and Archibald 2008). In any case, the limited resolution of the Pf and Pf a trees precludes us from making definitive conclusions. However, as the two proteins are tightly functionally linked, it is reasonable to suppose that they could share the same phylogenetic history. If so, then the support for a firmicute-Bacteroides relationship over α-proteobacterial or cyanobacterial affinity could be better assessed by concatenated analyses of these proteins.

Concatenation of Pf a and Pf Increases Monophyly Support and Suggests a Firmicute Ancestry for these Genes

We first tested the assumption that the two proteins shared the same history. In order to do this, all paralogous sequences were removed because some species had multiple Pf sequences but only one Pf a and thus each Pf could not be uniquely assigned a Pf a without duplication of the sequences. Congruence of the two protein phylogenies was tested using a likelihood ratio test implemented in CONCATERPILLAR (Leigh et al. 2008). Congruence was rejected (P value = 0.003657) with the full data sets with paralogs removed. We suspected that this may have been the result of different topologies observed within the δ-proteobacterial clade for the two proteins. As the latter group was strongly recovered as monophyletic (BP = 98 and PP = 1.0) and showed no affinity to the eukaryotic taxa in the phylogeny, we removed these sequences from the analysis and retested congruence; this time the null hypothesis of congruence could not be rejected (P = 0.056334). ML and BI were conducted on this reduced concatenated Pf–Pf a data set. For ML analyses, both partitioned (i.e., allowing gene-specific branch lengths and alpha shape parameters) and nonpartitioned analyses were performed. As differences between the bootstrap values generated from these two analyses were small (i.e., within Monte Carlo sampling error), we report only the unpartitioned bootstrap values. The eukaryotic sequences remain monophyletic with borderline significant corrected bootstrap support (BP = 95), but with surprisingly poor posterior probability (PP = 0.5) (fig. 2A). Similar to the Pf analysis (supplementary fig. S1, Supplementary Material online), the concatenated phylogeny shows a eukaryote–firmicute–Bacteroides grouping with strong support values (BP = 84, BPc = 96, PP = 1.0). All alternative topologies were rejected by AU tests (table 1), including cyanobacteria + eukaryote monophyly (Punpart = 0.00005 and Ppart = 0.009), which was previously not rejected.

Because some of the eukaryote Pf and/or Pf a sequences in our analyses were partial with small numbers of sites, we evaluated the impact of removal of taxa with less than 50% of the sites (fig. 2B). The tree topology did not change; however, support values for both eukaryotic monophyly (BP = 94, BPc = 98, and PP = 0.93) and firmicute–Bacteroides affinity (BP = 89, BPc = 97, and PP = 1.0) increased (fig. 2B). This observation suggests that Pf and Pf a within the eukaryotes had a single common prokaryotic origin probably resulting from a Lateral gene transfer (LGT) event from either a firmicute- or Bacteroides-like organism. Interpretation of the source of transfer is complicated by the scrambled bacterial relationships observed. Because bacterial “phyla” such as firmicutes, proteobacteria, and Bacteroides are not recovered as single groups but appear in multiple distinct
regions of the tree and the presence/absence of Pf or Pf homologs within bacterial genomes is patchily distributed across bacterial phyla, it is likely that LGT of these genes is occurring amongst these bacterial groups. Nevertheless, the most parsimonious interpretation of the placement of the eukaryotic clade within the bacteria suggests a firmicute ancestry.

Atypical Eukaryotic Relationships in Pf and Pf Phylogenies Suggest Eukaryote-to-Eukaryote LGT Events

The lack of evidence for a mitochondrial or cyanobacterial origin combined with apparent eukaryote monophyly in Pf and Pf phylogenies can be explained most straightforwardly by three hypotheses. First, the observed pattern

---

**Fig. 2.** Phylogeny of concatenated pyruvate formate lyase and its activating enzyme. The topology shown was obtained by ML analysis. Support values for branches are shown above them in the order of bootstrap support (BP)/corrected bootstrap support (BPc) and Bayesian posterior probability. Support values are only shown if with greater than 50% or 0.5. (A) A total of 887 sites from 63 taxa were used for this analysis. The conservation of operon architecture in bacteria or gene order in eukaryotes is shown in boxes to the right of the taxa where black and grey denote Pf and Pf respectively. The dotted boxes represent those taxa that are shown as a wedge in B. (B) The same analysis except removing eukaryotic taxa with fewer than 50% of the sites in the total alignment. A total of 887 sites from 54 taxa were used for this analysis.
could be explained by several independent prokaryote-to-eukaryote LGT events from a common or closely related prokaryotic source. Alternatively, there could have been a single origin of these enzymes by bacteria-to-eukaryote LGT along the lineage leading to the last common ancestral eukaryote followed by differential loss of the enzymes in disparate eukaryotic lineages (but retention in some facultative and obligate anaerobes). Finally, the enzymes could have been transferred from bacteria more recently to an extant eukaryote lineage and subsequently spread to other lineages by multiple eukaryote-to-eukaryote LGT events.

If the first scenario—the independent acquisition of Pf1/Pf1a from similar firmicute-like sources—were correct, then we would expect the Pf1 and Pf1a phylogenies to resolve typical eukaryotic relationships between organisms possessing these enzymes. To test this hypothesis, ML (partitioned and nonpartitioned) and BI analyses were performed on the concatenated twenty-one eukaryotic Pf1 and Pf1a sequences available (fig. 3). Again, the differences in BPs observed between partitioned and nonpartitioned analyses were within Monte Carlo error and so only the latter are shown.

The most striking observation from these analyses is the robust recovery of both the chlorophyte and prasinophyte green algal clades as branching strongly apart (in terms of high BP and PP values) in the tree, with all of the remaining eukaryotic lineages emerging between them. For example, the sequences from the amoebozoan Ma. balamuthi group with the chlorophytes to the exclusion of all other eukaryotes with significant support (BP = 95, BPc = 99, and PP = 1.0). Another odd feature of the analyses is the clustering of the ichthyosporean Am. parasiticum with the glaucophytes and rhodophytes and not the other opisthokonts (Neocallimastix frontalis and Piromyces sp.).
The placement of the glaucophyte and rhodophytes in relation to the other lineages is not well resolved which has been observed previously in multiprotein eukaryotic phylogenies (Hampl et al. 2009; Nozaki et al. 2009). Finally, the haptophyte and diatoms branch together with strong support (BP = 100, BC = 100, and PP = 1.0).

To examine whether these unexpected eukaryotic relationships reflected a strong signal in the data versus random error, topology tests were performed in both unpartitioned and partitioned analyses (table 1 and schematic representations of topologies given in Supplementary fig. S3, Supplementary Material online). We first tested the compatibility of the data with a conservative view of known eukaryotic relationships. The following groups were enforced as topological constraints during the tree searching: monophyly of the unikonts with the amoebozan Mastigamoeba to the immediate exclusion of monophyletic opisthokonts (chytrids and the ichthyosporean) and monophyly of the bikonts with the haptophyte and stramenopile (diatom) to the immediate exclusion of the archaeoplastid groups (prasinophytes, chlorophytes, rhodophytes, and glaucophytes). The optimal trees estimated with this “typical eukaryotic phylogeny” constraint were rejected in both nonpartitioned and partitioned analyses (Punpart = 0.002) and partitioned analyses (Ppart = 0.003). The monophyly of the green algal lineages (prasinophytes and chlorophytes) was also rejected for both analyses (Punpart = 0.01 and Ppart = 0.008) as was monophyly of the green algal lineages plus Ma. balamuthi (Punpart = 0.038 and Ppart = 0.0330) suggesting that it is not only the placement of Ma. balamuthi that is responsible for the separation of the algae. Finally, opisthokont monophyly (i.e., Am. parasiticum + chytrid fungi) was tested but not rejected (Punpart = 0.6610 and Ppart = 0.6650). Thus, only in the latter case could random error alone be responsible for the anomalous branching pattern. Addition of more taxa may be helpful to improve resolution of this region of the phylogeny.

The foregoing analyses suggest that many of the peculiar eukaryotic relationships observed in the tree derive from strong signals in the data. To test whether the observed relationships were phylogenetic artifacts, we analyzed the amino acid compositions of each of the eukaryotic taxa relative to the full data set with a chi-squared test implemented in TREE-PUZZLE (Schmidt et al. 2002). We found significant deviation in the amino acid compositions of the sequences from Mi. pusilla (P = 0.0001 Pprl = 0.00384) and Neocallimastix (P = 0.0036). Reanalyses of the data sets after removing these two taxa from the concatenated analysis did not significantly change the overall tree topology or support values for strongly supported branches (data not shown). Furthermore, bootstrapped clustering analyses of the amino acid composition vectors of sequences in the eukaryotic and concatenated data set showed no strongly supported internal structure. The deepest clusters recovered in these dendrograms did not resemble deep groupings in ML or Bayesian phylogenies suggesting that amino acid composition alone is not biasing the estimated phylogenies (data not shown). Finally, removal of fast evolving sites using the AIR package (Kumar et al. 2009) did not change the estimated topology, although, not surprisingly, support for eukaryotic monophyly and firmicute affinity generally decreased after removal of the fastest-evolving ~50% of the positions from the alignment (supplementary fig. S4, Supplementary Material online).

Therefore, in the absence of any recognizable source of systematic error in these data sets (e.g., long-branch attraction or amino acid composition bias), the atypical eukaryotic relationships observed likely reflect a true historical signal. This, combined with the extreme patchiness of Pr and Pf genes observed across eukaryotic diversity (fig. 1), disfavors the “ancestral plus differential loss” scenario. A variation of this scenario that invokes large numbers of early gene duplications within eukaryotes followed by differential loss is conceivable and would explain the odd phylogenetic patterns observed. As none of the eukaryotes so far examined have retained multiple putative “ancient paralogs” expected under this scenario and as the number of parallel loss events required to explain the extant presence/absence pattern becomes even greater as does the number of paralogs that had to have been in the common eukaryotic ancestor, it seems to be a rather unlikely explanation. This leaves the hypothesis of a firmicute-to-eukaryote LGT event into an extant eukaryote lineage followed by several subsequent eukaryote-to-eukaryote gene transfers as a remaining reasonable scenario. The number of documented LGTs from prokaryotes to eukaryotes has increased steadily as more genomic data becomes available (see Andersson 2009 for a recent review). Similarly, the frequency of reports of eukaryote-to-eukaryote LGT events have also increased; this process may occur via mechanisms, such as secondary plastid symbioses (Kamikawa et al. 2009), simple phagotrophy (Andersson 2009), or via parasitic genetic elements (Gilbert et al. 2010). Therefore, a “multiple LGT” scenario for Pr/Pf evolution within eukaryotes seems to be at least a plausible explanation for the phylogenetic patterns we have observed.

Assuming, for the moment, that the foregoing LGT scenario for Pf and Pf were correct, it is possible to make informed conjectures about the sequence of events that took place. For example, these genes could have been acquired by an ancestral archaeoplastid organism and were directly inherited by Cyano. paradoxa, Por. haitanensis and the two green algal lineages. Assuming that these chlorophytes and prasinophytes form a clade to the exclusion of streptophytes (Baurain et al. 2010) then loss of Pf and Pf would have to be invoked in the latter lineage and in the red alga Cyanidioschyzon merolae (fig. 3). The remainder of eukaryotic organisms in fig. 3 could have acquired these genes then by a minimum of four distinct eukaryote-to-eukaryote LGT events from the various archaeoplastid lineages. However, many other alternative scenarios that invoke a similar number of LGT events from different founding eukaryotic lineages are also possible. Ultimately, much more genomic data from many more representatives of the various microbial eukaryotic taxa involved
will be needed to 1) confirm or refute the general LGT scenario we have outlined and 2) clarify the timing of the various genetic events involved in the evolution of Pfl/Pfla within eukaryotes.

Conservation of Gene Order: Insights into the Transfer Mechanism

One potential difficulty with an LGT origin scenario for Pfl and Pfla genes is that for a functioning enzyme system to be “moved” between organisms, at a minimum, both genes must be transferred successfully. However, for the initial bacteria-to-eukaryote transfer, this problem is avoided if the two genes were genetically linked in an operon that could be transferred in a single genetic event (Andersson and Roger 2003). In γ-proteobacteria, such as Escherichia coli and most firmicutes, pfl and pfla are indeed encoded close together in a putative operon as predicted by OperonDb (Pertea et al. 2009). A survey of the gene organization and order is summarized in figure 2 for those taxa analyzed in the concatenated gene analysis. Separation of pfl and pfla loci in bacteria was only observed in the Streptococcaceae lineage (Lactococcus and Streptococcus) and Thermosynechococcus elongates (fig. 2); neither of these groups is closely related to the eukaryotic clade. The initial LGT into eukaryotes plausibly occurred from a firmicute-like groups is closely related to the eukaryotic clade. The initial genetic events involved in the evolution of Pfl/Pfla within eukaryotes.

OperonDb (Pertea et al. 2009). A survey of the gene

coli and most firmicutes, pfl and pfla are indeed encoded close together in a putative operon as predicted by OperonDb (Pertea et al. 2009). A survey of the gene organization and order is summarized in figure 2 for those taxa analyzed in the concatenated gene analysis. Separation of pfl and pfla loci in bacteria was only observed in the Streptococcaceae lineage (Lactococcus and Streptococcus) and Thermosynechococcus elongates (fig. 2); neither of these groups is closely related to the eukaryotic clade. The initial LGT into eukaryotes plausibly occurred from a firmicute-like groups is closely related to the eukaryotic clade. The initial genetic events involved in the evolution of Pfl/Pfla within eukaryotes.

Gene Structures and Organization of Pfl Genes in Diverse Eukaryotes

No obvious sequence and little positional conservation of introns was observed across lineages after scanning genomic data from Os. tauri, Os. lucimarinus, Mi. pusilla, Micromonas sp., Tha. pseudonana, Chla. reinhardtii, Piromyces sp., Chlorella sp., and V. carteri. The Ostreococcus species did not contain any introns in pfl or pfla genes. Within the Micromonas species, there was intron position conservation in pfla however the sizes of the introns differed significantly (72 bp in Mi. pusilla and 293 bp in Micromonas sp.). This might represent an ancestral intron of the Micromonas lineage that was acquired after speciation from the other prasinophyte lineage (Ostreococcus) or, alternatively, it could be older but was lost in the Ostreococcus lineage.

In Silico Predictions of Subcellular Localization

Previous reports have localized Pfl activity to the hydrogenosomes of Neo. frontalis and Piromyces sp. and to the chloroplasts and mitochondria of Chla. reinhardtii (Boxma et al. 2004; Celius-Dietrich and Henze 2004; Atteia et al. 2006; Hemschemeier et al. 2008). To elucidate the organellar localization of the newly identified eukaryotic Pfl and Pfla sequences, we utilized the software tools TargetP and Mitoprot II. In all cases, except one, these programs did not strongly predict an organellar localization of Pfl or Pfla in the newly described sequences. The exception was a “weakly” predicted mitochondrial targeting peptide in the Pfla of Os. tauri (Ptarget = 0.319 and Pmitoprot = 0.773). It should be noted that all of the previously characterized organellar homologs were in fact predicted to possess mitochondrial targeting peptides by one or both of these programs. Therefore, at this stage, there seems no evidence for an organellar localization of the newly described Pfl and Pfla homologs; they may in fact all function within the cytoplasm. Obviously, further molecular and biochemical experiments are needed to confirm or refute a cytosolic localization as “in silico” localization prediction has limited accuracy on nonmodel organisms and many organellar proteins possess cryptic targeting signals.

The Distribution of Pyruvate Catalyzing Enzymes in Eukaryotes

The discovery of enzymatic alternatives to Pdh, such as Pfo, Pno, and Pfl in some eukaryotes that all generate acetyl-
CoA raises the question of their *raison d’être*. Some organisms such as *Chlamydomonas* and *Thalassiosira* appear to contain Pfo, Pfl, and Pdh (fig. 1) which may indicate differential compartmentalization of these acetyl-CoA generating enzymes within these organisms. In *Chlamydomonas*, Pfl functions in both the chloroplast and the mitochondrion under anaerobic conditions suggesting that this acetyl-CoA generating enzyme might functionally replace Pdh under anaerobic conditions or under conditions where high CO2 or redox levels are inhibiting Pdh- or Pfo-mediated catalysis. This hypothesis is supported by the fact that Pfl functions in the absence of redox components typically associated with acetyl-CoA generation, such as NAD(P)+ or ferredoxin. In fact, activation of Pfl by Pfala likely requires reduced components—more specifically ferredoxin—in species of *Clostridium* (Thauer et al. 1972 1972). Therefore, the main need for Pfl may arise under anaerobic conditions where reduced cofactors/proteins are at high levels, but acetyl-CoA is still needed for anabolic pathways, such as lipid and amino acid biosynthesis or even ATP generation. As mentioned earlier, initial Pfl activation is dependent on SAM whose synthesis requires ATP, therefore, the radical must be protected in order to not lose this energy investment.

**Conclusions**

Our survey of eukaryotic genomic data revealed a wide diversity of microbial eukaryotic lineages that possess Pfl homologues and their activating enzymes. Many of these organisms (e.g., the diatom *Tha. pseudonana*, the seaweed *Por. haitanensis*, and the glaucophyte *Cyano. paradoxa*) are traditionally thought of as “aerobic” eukaryotes, but it is likely that they also transiently experience low oxygen conditions and this oxygen-sensitive system may then be important for continued production of acetyl-CoA. We have shown that all Pfl homologues and their activating enzymes encoded in eukaryote genomes form a monophyletic group that were likely acquired once from a bacterial source. The eukaryotic enzymes appear to be most closely related to homologues found in selected firmicute gram positive bacteria and show no evidence of a mitochondrial or chloroplast endosymbiotic ancestry. The most parsimonious explanation for eukaryote monophyly in this case is a single LGT event into a eukaryotic lineage, probably from a firmicute donor. The patchy distribution of Pfl and Pfala across eukaryotic diversity and strongly supported atypical eukaryotic relationships recovered in the phylogenies of these enzymes argues against an origin of the enzymes in the common ancestral eukaryotic lineage followed by differential lineage-specific loss. Rather, we propose that after the initial transfer of the two genes encoded in a bacterial operon into a eukaryotic lineage, subsequent eukaryote-to-eukaryote transfers of the two genes occurred giving rise to the present-day Pfl and Pfala distribution in eukaryotes. This scenario is made more plausible by our finding that the two genes are sometimes closely genetically linked (or in one case fused) in eukaryotic genomes so that single transfer events may be sufficient to spread a functional Pfl system between eukaryotes. Although we favor the latter scenario, it should be regarded as tentative as a denser and broader sampling of eukaryotic microbial genomes will be required to confirm, refine, or refute it.

The probable spread of Pfl and Pfala via LGT within the eukaryotic realm questions the assumption, often made by comparative genomic analyses, that orthologues detected in a few distantly related eukaryotes can be automatically assumed to have been present in the common ancestor of all eukaryotes. Additionally, our reconstruction of the evolutionary histories of Pfl and Pfala does not straightforwardly fit with the scenario for the origin of eukaryotic enzymes of anaerobic metabolism advanced in the Hydrogen Hypothesis (Martin and Müller 1998). A simple interpretation of the Hydrogen Hypothesis would predict that these enzymes should be present in the eukaryotic common ancestor and should derive from the mitochondrial endosymbiont. Neither of these claims are supported by our analyses of the Pfl and Pfala data.

Finally, although the rationale for the presence of multiple pyruvate catalyzing enzymes in some eukaryotes remains uncertain, it likely involves the differential regulation of these enzymes under different oxygen tension and redox conditions. Experimental testing of gene expression, enzyme activities, and subcellular localization of these proteins under various conditions in the organisms that possess them will be required to fully understand their biochemical roles and evolutionary histories.

**Supplementary Material**

Supplementary table S1 and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We thank Jacque de Mestral for technical assistance; Dan Gaston, Martin Kolisko, and Ed Susko for help and advice with analyses; and Michelle Leger for critical comments on the manuscript. This work was supported by Canadian Institutes for Health Research (grant number MOP 62809 awarded to A.J.R.). C.W.S. is supported by scholarships from the Natural Sciences and Engineering Research Council of Canada and Killiam Trusts. V.H. was supported by Ministry of Education, Youth and Sport of the Czech Republic (project MSM0021620828).

**References**


