Research Article

Journal of
Molecular Microbiology
and Biotechnology

J Mol Microbiol Biotechnol 2003;5:172–189
DOI: 10.1159/000070268

Phylogenomic Analysis of the *Giardia intestinalis* Transcarboxylase Reveals Multiple Instances of Domain Fusion and Fission in the Evolution of Biotin-Dependent Enzymes

I. King Jordan[a] Katrin Henze[b,c] Natalie D. Fedorova[a] Eugene V. Koonin[a] Michael Y. Galperin[a]

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Md., and [b]The Rockefeller University, New York, N.Y., USA; [c]Institute of Botany III, Heinrich Heine University of Düsseldorf, Düsseldorf, Germany

Key Words
Pyruvate carboxylase · Oxaloacetate decarboxylase · Acetyl-CoA carboxylase · Biotin · Domain fusion · Phylogenetic tree

Abstract
Sequencing of the gene encoding a pyruvate carboxylase-like protein from the amitochondrial eukaryote *Giardia intestinalis* revealed a 1,338 aa protein composed of acetyl-CoA carboxyltransferase (ACCT), pyruvate carboxyltransferase (PycB), and biotin carboxyl carrier protein (BCCP) domains, linked in a single polypeptide chain. This particular domain combination has been previously seen only in the methylmalonyl-CoA:pyruvate transcarboxylase from *Propionibacterium freudenreichii*, where each of these domains is encoded by an individual gene and forms a separate subunit. To get an insight into the evolutionary origin and biochemical function of the *G. intestinalis* enzyme, we compared its domain composition to those of other biotin-dependent enzymes and performed a phylogenetic analysis of each of its domains. The results obtained indicate that: (1) evolution of the BCCP domain included several domain fusion events, leading to the ACCT-BCCP and PycB-BCCP domain combinations; (2) fusions of the PycB and BCCP domains in pyruvate carboxylases and oxaloacetate decarboxylases occurred on several independent occasions in different prokaryotic lineages, probably due to selective pressure towards co-expression of these genes, and (3) because newly sequenced biotin-dependent enzymes are often misannotated in sequence databases, their annotation as either carboxylases, decarboxylases, or transcarboxylases has to rely on detailed analysis of their domain composition, operon organization of the corresponding genes, gene content in the particular genome, and phylogenetic analysis.

Copyright © 2003 S. Karger AG, Basel

The first two authors contributed equally to this paper.

Introduction

Giardia intestinalis, also known as *Giardia lamblia* or *Giardia duodenalis* [see Kulda and Nohýnková, 1995], is a microaerophilic intestinal parasite that belongs to the group of amitochondriate eukaryotes without metabolic compartmentalization (Type I amitochondriates) [Kulda and Nohýnková, 1995; Martin and Müller, 1998; Müller,

KARGER
Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2003 S. Karger AG, Basel
1464–1801/03/0053–0172$19.50/0

Accessible online at:
www.karger.com/mmb

Michael Y. Galperin
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894 (USA)
Tel. +1 301 435 5910, Fax +1 301 435 7794, E-Mail galperin@ncbi.nlm.nih.gov

Material may be protected by copyright law (Title 17, U.S. Code)

1998; Reeves, 1984; Sogin et al., 1989]. Recent studies indicate that Type I amitochondriates descend from mitochondrion-containing ancestors by secondary structural and functional losses [Roger, 1999]. Studies of these early-diverging organisms provide insights into the evolution of eukaryotic cells [Embley and Hirt, 1998; Hilario and Gogarten, 1998; Morrison et al., 2001; Sogin et al., 1989; Triana et al., 2001] and hold promise in developing new therapies against the diseases caused by protozoan pathogens.

It has been well documented that protozoan genomes encode a complex mix of proteins, some of which seem to have bacterial or archaeal origin, whereas others are most closely related to metazoan orthologs. Indeed, fermentation enzymes of *G. intestinalis* and *Entamoeba histolytica* seem to have evolved from bacterial ancestors and could have had protomitochondrial origin [Brown et al., 1998; Henze et al., 1998; Rosenthal et al., 1997]. In contrast, acetyl-CoA synthetase (ADP-forming) found in both of these organisms, as well as GTP-dependent phosphoenolpyruvate carboxykinase of *G. intestinalis* appear to have archaeal origin [Field et al., 2000; Sanchez et al., 2000; Suguri et al., 2001]. Naturally, many protozoan systems are specifically eukaryotic, or, like histones [Wu et al., 2000] and proteasome [Bouzat et al., 2000], are much closer to their metazoan counterparts than to archaeal ones.

The core metabolism of *Giardia*, as well as of other Type I amitochondriates, is cytosolic [Martin and Müller, 1998; Müller, 1998; Reeves, 1984]. The principal energy substrate is glucose, which is catabolized via the classical Embden-Meyerhof-Parnas glycolytic pathway with acetate and ethanol as the main metabolic end products [Müller, 1998; Reeves, 1984]. Tricarboxylic acid cycle enzymes and cytochrome-mediated electron transport chain components appear to be missing in *Giardia*, and substrate level phosphorylation of ADP or GDP by pyruvate kinase and acetyl-CoA synthetase has an important role in its energy metabolism [Park et al., 1997; Sanchez et al., 1999; Sanchez et al., 2000].

The absence of mitochondria leaves the cytoplasmic membrane as the only energy-transducing membrane in *Giardia*. Although *Giardia* appears to have no respiratory ion pumps that would generate a transmembrane gradient of H^+ (proton-motive force), it has been reported to maintain a membrane potential of −134 mV [Biagini et al., 2000], probably through the action of the membrane-bound V-type H^+-ATPase [Hilario and Gogarten, 1998]. In addition, *Giardia* has been shown to maintain its cytoplasmic pH in a Na^+-dependent manner, raising the possibility that, like many bacterial pathogens [Häse et al., 2001], it could use transmembrane gradient of Na^+ ions (sodium-motive force) in its energy metabolism. Although no primary Na^+ pumps, such as Na^+-translocating NADH:ubiquinone oxidoreductase, oxaloacetate decarboxylase, or coenzyme M methyltransferase [Dimroth, 1997; Häse et al., 2001; Hayashi et al., 2001], have been found in *Giardia* so far, an open reading frame with some similarity to oxaloacetate decarboxylase has been discovered recently as part of the *Giardia* genome sequencing project undertaken at the Woods Hole Marine Laboratory [McArthur et al., 2000] and annotated as pyruvate carboxylase. Because *Giardia* had been shown to possess a pyruvate kinase, a pyruvate, phosphate dikinase, and a PEP carboxykinase [Bruderer et al., 1996; Lindmark, 1980; Nevalainen et al., 1996; Park et al., 1997; Suguri et al., 2001], it was not immediately clear what would be the metabolic role of a pyruvate carboxylase in this organism. However, there was an attractive possibility that this gene could encode an oxaloacetate decarboxylase, a primary Na^+ pump. To characterize this novel open reading frame and investigate the possibility that it encodes an oxaloacetate decarboxylase, we completely sequenced the corresponding gene and identified its product. Here we report sequence analysis of this enzyme, which revealed an unusual domain composition, seen previously only in the methylmalonyl-CoA:pyruvate transcarboxylase from *Propionibacterium freudenreichii* subsp. *shermanii* [Wood, 1979; Wood and Kumar, 1985]. In contrast to the *P. freudenreichii* transcarboxylase, the *G. intestinalis* enzyme contained the respective domains in a single polypeptide chain. Phylogenetic analysis of these domains and a comparison of the domain architectures of various biotin-dependent enzymes indicated that evolution of biotin-dependent enzymes involved several independent instances of domain fusion and probably also fission. These results should be useful in assigning functions to other biotin-dependent enzymes identified in genome-sequencing projects.

Results

Domain Organization of the Giardia *Protein and Other Biotin-Dependent Enzymes*

Sequence database searches with the 1,338-aa open reading frame from *G. intestinalis* as a query revealed a biotin-dependent enzyme with a unique domain organization. According to the BLAST results, its N-terminal region showed significant sequence similarity to bacterial

Material may be protected by copyright law (Title 17, U.S. Code)

and eukaryotic acetyl-CoA and propionyl-CoA carboxylases, whereas the C-terminal region showed strongest similarity to pyruvate carboxylases from gram-positive bacteria, such as *Bacillus subtilis* and *Bacillus stearothermophilus* [Kondo et al., 1997]. It was clear therefore that this *Giardia* protein was composed of several distinct domains with different phylogenetic affinities. With multidomain organization of the query protein and/or its database hits being a major problem in interpreting the BLAST results [Galperin and Koonin, 1998], assignment of a potential function to this *Giardia* protein had to be based on a detailed analysis of the domain organization of this protein and other biotin-dependent enzymes.

Analysis of the subunit and domain organization of various biotin-dependent enzymes (fig. 1) showed that their only common feature was the presence of the biotin carboxyl carrier protein (BCCP) subunit (or domain). Although all biotin-dependent enzymes contain a carboxyltransferase domain, sequence analysis identifies two distinct classes of this domain. Acetyl-CoA and propionyl-CoA carboxyltransferase domains (ACCT) are homologous, but are unrelated to the pyruvate carboxyltransferase (PycB) domains, found in pyruvate carboxylases and oxaloacetate decarboxylase. Each ACCT domain (or subunit) consists of two subdomains (or subunits), AccA and AccD, which are distantly related to each other. In addition, enzymes involved in CO_2 assimilation (carboxylases) contain a biotin carboxylase (BC) subunit or domain, whereas biotin-dependent decarboxylases contain a membrane-bound decarboxylase (OadB) subunit. Methylmalonyl-CoA:pyruvate transcarboxylase from *P. freudenreichii* subsp. *shermanii* (formerly *P. shermanii*), the only enzyme of this class described so far, contains a BCCP subunit and two different carboxyltransferase subunits, ACCT and PycB (fig. 1).

All biotin-dependent enzymes share a similar catalytic mechanism, which involves two partial reactions: (1) carboxylation of the biotin moiety, covalently attached to the BCCP, and (2) carboxyl transfer from biotin to the acceptor molecule [see Cronan and Waldrop, 2002, for a recent review]. The range of reactions catalyzed by each of the subunits can be described as follows:

Biotin carboxylase:

$$HCO_3^- + ATP + BCCP \leftrightarrow {}^-OOC\text{-}BCCP + ADP + P_i \quad (1)$$

AccA + AccD-type carboxyltransferase:

$$^-OOC\text{-}BCCP + \text{acetyl-CoA} \leftrightarrow BCCP + \text{malonyl-CoA} \quad (2)$$

$$^-OOC\text{-}BCCP + \text{propionyl-CoA} \leftrightarrow BCCP + \text{methylmalonyl-CoA} \quad (3)$$

$$^-OOC\text{-}BCCP + \text{crotonyl-CoA} \leftrightarrow BCCP + \text{glutaconyl-CoA} \quad (4)$$

$$^-OOC\text{-}BCCP + \text{methylcrotonyl-CoA} \leftrightarrow BCCP + \text{methylglutaconyl-CoA} \quad (5)$$

PycB/OadA-type carboxyltransferase:

$$^-OOC\text{-}BCCP + \text{pyruvate} \leftrightarrow BCCP + \text{oxaloacetate} \quad (6)$$

OadB subunit of membrane-bound decarboxylases:

$$^-OOC\text{-}BCCP \rightarrow BCCP + HCO_3^- + \Delta\mu_{Na^+} \quad (7)$$

where $\Delta\mu_{Na^+}$ indicates transmembrane electrochemical gradient of Na^+ ions (sodium-motive force). Biotin-dependent enzymes with various combinations of these subunits (fig. 1) have been shown to catalyze the following overall reactions: carboxylation of acetyl-CoA [(1) + (2)], propionyl-CoA [(1) + (3)], crotonyl-CoA [(1) + (4)], or methylcrotonyl-CoA [(1) + (5)]; pyruvate carboxylation [(1) + (6)]; decarboxylation of methylmalonyl-CoA [(3) + (7)], glutaconyl-CoA [(4) + (7)], and oxaloacetate [(6) + (7)]; and transcarboxylation of methylmalonyl-CoA and pyruvate [(3) – (6)].

The *Giardia* protein turned out to be composed of two distinct carboxylase domains, followed by the BCCP domain at its very carboxy terminus (fig. 1). Judging from its domain composition, the *Giardia* enzyme is expected to couple reaction (6) with reactions (2), (3), (4), or (5), i.e. catalyze carboxyl transfer from oxaloacetate to an acyl-CoA or vice versa. The only previously characterized enzyme with a similar domain composition (fig. 1) is the methylmalonyl-CoA:pyruvate transcarboxylase from *P. freudenreichii* [Wood, 1979; Wood and Kumar, 1985]. However, in the *P. freudenreichii* enzyme, acetyl-CoA carboxyltransferase domain, pyruvate carboxyltransferase domain, and BCCP are represented by separate subunits (12S, 5S, and 1.3S subunit, respectively) [Wood, 1979; Wood and Kumar, 1985]. The *Giardia* enzyme thus has a previously unknown multidomain architecture. Indeed, while fusions of BCCP and the pyruvate carboxyltransferase domain are common in pyruvate carboxylases and oxaloacetate decarboxylases (fig. 1), this is the first instance of a fusion between an acetyl-CoA carboxyltransferase domain and a pyruvate carboxyltransferase domain.

A sequence database search for other multidomain proteins with the same AccD-AccA-PycB domain fusion revealed an additional case of such fusion encoded in the unfinished genome sequence of another protist, *E. histolytica*, available at the TIGR web site [http://www.tigr.org/]. Its closest database hits, other than the *Giardia* protein, were also among the enzymes from gram-positive bacteria (data not shown). To better understand the evolutionary origin and possible functions of the novel fusion enzyme detected in *G. intestinalis* and *E. histolytica*, we examined the phyletic distribution, gene neighborhood, and evolutionary history of its constituent domains.

Material may be protected by copyright law (Title 17, U.S. Code)

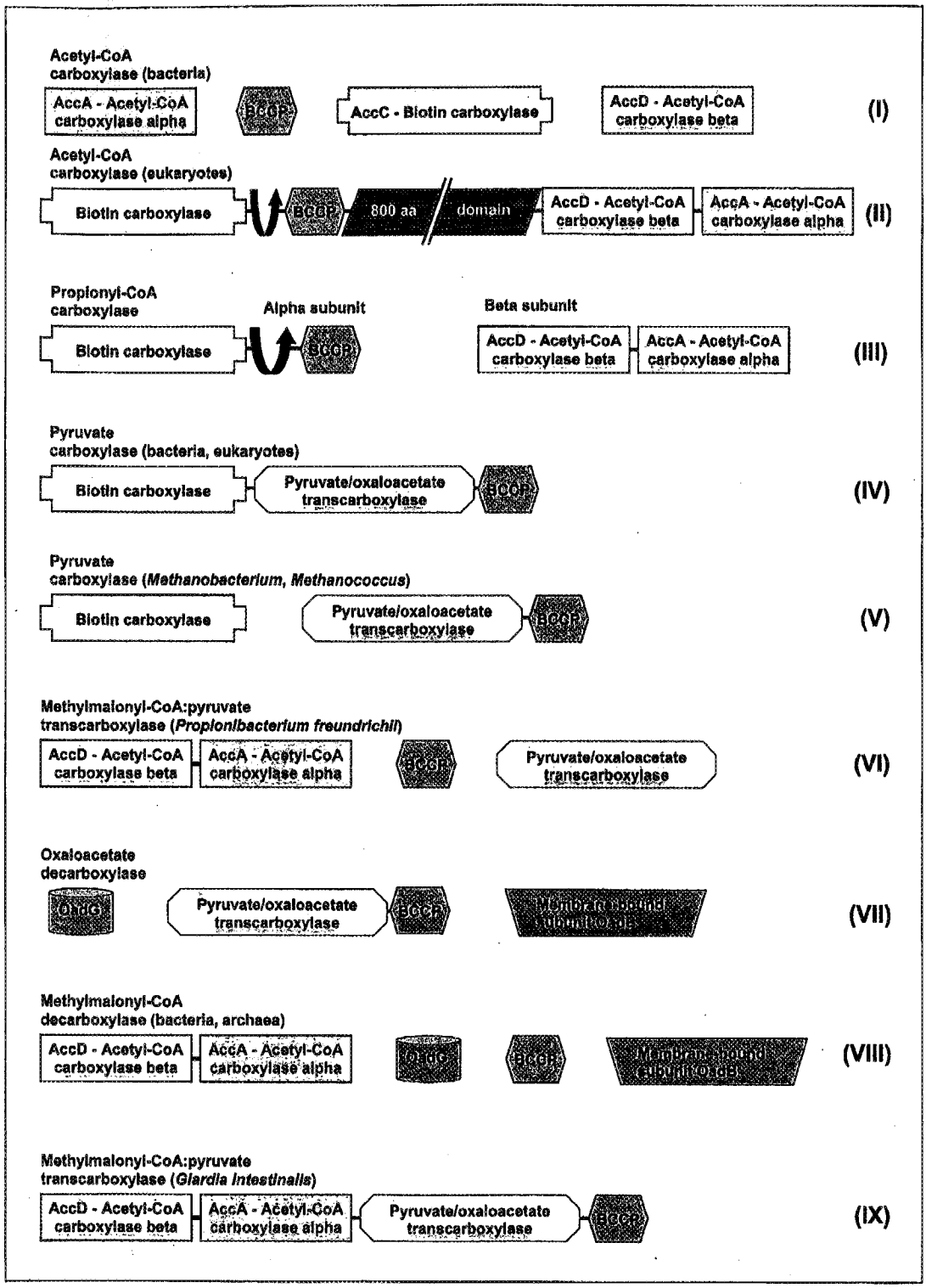


Fig. 1. Domain organization of biotin-dependent enzymes. Each shape indicates a specific conserved domain and its approximate length. *accA*- and *accD*-encoded α and β subunits of the *Escherichia coli* acetyl-CoA carboxylase are both shown as rectangles to indicate that they are paralogs. Fused domains that form a single polypeptide chain are indicated by a connecting line.

Material may be protected by copyright law (Title 17, U.S. Code)

Table 1. Genomic context of the biotin-dependent enzymes

Enzyme	Acetyl-CoA carboxylase				Oxaloacetate decarboxylase		
			Pyruvate carboxylase				
Domain name	Acetyl-CoA carboxyl-transferase		Biotin carboxylase	BCCP	Pyruvate carboxyl-transferase	Beta subunit	Gamma subunit
Gene name[a]	AccD	AccA	AccC	AccB	PycB/OadA	OadB	OadG
COG number	0777	0825	0439	0511	1038/3632	1883	3630
Archaea							
Aeropyrum pernix	–	–	–	–	–	–	–
Archaeoglobus fulgidus	AF2217	AF2217	AF0220	AF2216 AF2085	AF1252	AF2084	AF2086
Halobacterium sp.	Vng0623g Vng1529g	Vng0623g Vng1529g	Vng1532g	Vng1532g		–	–
Methanococcus jannaschii	–	–	MJ1229	MJ1231	MJ1231	–	–
Methanobacterium thermoautotrophicum	–	–	MTH1917	MTH1107	MTH1107	–	–
Pyrococcus abyssi	PAB1769	PAB1769	–	PAB1771	–	PAB1772	PAB1770
Pyrococcus horikoshii	PH1287	PH1287	–	PH1284 PH0834	PH0834	PH1282	PH1283
Sulfolobus solfataricus	SSO2463	SSO2463	SSO2466	SSO2464	–	–	–
Thermoplasma acidophilum	–	–	–	–	–	–	–
Bacteria							
Aquifex aeolicus	aq_445	aq_1206	aq_1517 aq_1664 aq_1470	aq_1520 aq_1514 aq_1363	aq_1520 aq_1514	–	–
Thermotoga maritima	TM0716	TM0716	–	TM0717	TM0128	TM0880	?
Synechocystis sp.	sll0336	sll0728	sll0053	slr0435	–	–	–
Chlamydia trachomatis	CT293	CT265	CT124	CT123	–	–	–
*Deinococcus radiodurans**	DRA0310 DR1616	DRA0310 DR1616	DRA0310 DR0117	DRA0310 DR0118	–	–	–
Treponema pallidum	–	–	TP0695	TP0056	TP0056	TP0057	TP0055
Gram-positive bacteria							
Bacillus subtilis	BS_yqjD	BS_yqjD BS_accA	BS_pycA BS_accC	BS_pycA	BS_pycA	–	–
Lactococcus lactis	L0180	L0190	L0189 L63652	L0187 L63652	L63652	–	–
Streptococcus pyogenes	SPy1744	SPy1743	SPy1745 SPy0560	SPy1747 SPy1176	SPy1174	SPy1177	?
				SPy0183	SPy0182	SPy0184	?
*Mycobacterium tuberculosis**	Rv0904c Rv0974c Rv3280	Rv0904c Rv0974c Rv3280	Rv2967c Rv0973c Rv3285	Rv2967c Rv0973c Rv3283	Rv2967c	–	–
*Mycobacterium leprae**	ML0731	ML0731	ML0726	ML0726	–	–	–

Material may be protected by copyright law (Title 17, U.S. Code)

Table 1 (continued)

Enzyme	Acetyl-CoA carboxylase				Oxaloacetate decarboxylase		
			Pyruvate carboxylase				
Domain name	Acetyl-CoA carboxyl-transferase		Biotin carboxylase	BCCP	Pyruvate carboxyl-transferase	Beta subunit	Gamma subunit
Gene name[a]	AccD	AccA	AccC	AccB	PycB/OadA	OadB	OadG
COG number	0777	0825	0439	0511	1038/3632	1883	3630
Alpha-proteobacteria							
*Mesorhizobium loti**	mll0348	mll0348	mll3969	mll3969	mll3969		
*Caulobacter crescentus**	CC3214	CC3214	CC3214	CC3214	–	–	–
	CC1975	CC1975	CC1829	CC1829			
	CC2170	CC2170	CC2168	CC2168			
	CC3542	CC2995	CC2179	CC2179			
			CC1884	CC1883			
Rickettsia prowazekii	RP619	RP619	RP618	RP618	–	–	–
Gamma-proteobacteria							
Escherichia coli	accD	accA	accC	accB	–	–	–
Pseudomonas aeruginosa	PA1400	PA1400	PA1400	PA1400		–	–
	PA2014	PA3639	PA2012	PA2012			
	PA3436		PA3436	PA3436	PA3436		
Pasteurella multocida	PM0636	PM0292	PM1091	PM1092			
				PM1422	PM1422	PM1423	M1421
Vibrio cholerae	VC1000	VC2244	VC0295	VC0296			
				VC0550	VC0550	VC0551	VC0549
				VC0793m	VC0793m	VC0792	VC0792
Eukaryotes							
Yeast	YMR207c	YMR207c	YBR207c	YBR207c		–	–
	YNR016c	YNR016c	YNR016c	YNR016c			
			YBR218c	YBR218c	YBR218c		
			YGL062W	YGL062W	YGL062W		

[a] The gene names are taken from the complete genome sequences; they can be used to retrieve corresponding DNA sequences from GenBank or the deduced proteins from the NCBI protein database. Adjacent genes that form a (predicted) operon or are fused together and encode a multi-domain protein, are shown on the same line and are similarly shaded. For organisms indicated with asterisks, only a fraction of the encoded homologs are shown. The dash indicates the absence of the corresponding gene(s) in the complete genome. Question marks indicate still unidentified gamma subunits of oxaloacetate decarboxylase. Gene and operone assignments were made on the basis of the Clusters of Orthologous Groups of proteins (COG) database (http://www.ncbi.nlm.nih.gov/COG [Tatusov et al., 2000]). The frameshift mutation in the *V. cholerae* VC0793 gene was corrected and its product is referred to as VC0793m.

Operon Organization and Domain Fusions of Biotin-Dependent Enzymes

In bacteria and archaea, BCCP, biotin carboxylase and carboxyltransferase subunits that comprise acetyl-CoA carboxylase and propionyl-CoA carboxylase are usually encoded by adjacent but separate genes (table 1). In eukaryotes, these genes are usually fused and form large multi-domain proteins. Eukaryotic acetyl-CoA carboxylases, in contrast to propionyl-CoA carboxylases, contain an additional large (~800 aa) internal domain of unknown, presumably non-enzymatic, function that is inserted between the BCCP and ACCT domains (fig. 1). Pyruvate carboxylases from *B. subtilis*, several other bacteria, and yeasts are made up of biotin carboxylase (BC), pyruvate carbox-

Material may be protected by copyright law (Title 17, U.S. Code)

yltransferase (PycB), and BCCP domains joined in a single polypeptide chain [Morris et al., 1987; Mukhopadhyay and Purwantini, 2000; Sumper and Riepertinger, 1972]. In contrast, in the archaea *Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum*, only pyruvate carboxyltransferase and BCCP domains are fused, whereas biotin carboxylase forms a separate subunit [Mukhopadhyay et al., 1998; Mukhopadhyay et al., 2000]. The same PycB-BCCP fusion has been found in oxaloacetate decarboxylases and the decarboxylase subunits of similar enzymes described so far in several anaerobic bacteria and archaea [Buckel, 2001; Dimroth and Schink, 1998]. In contrast, no gene fusions have been reported for the subunits of biotin-dependent transcarboxylase [Wood, 1979; Wood and Kumar, 1985].

Genomic Context-Based Functional Assignments for Biotin-Dependent Enzymes

Comparative analysis of biotin-dependent enzymes encoded in completely sequenced bacterial, archaeal, and eukaryotic genomes revealed an extreme diversity of domain organization (table 1). In most cases, however, the enzymatic activity of the encoded protein could be deduced from its domain composition or, in case of separate genes, from the likely subunit composition determined by the genome content (i.e. the presence of only certain domains in a particular organism). Assuming that neighboring genes are co-regulated (belong to the same operon) and that their products interact with each other, most of the biotin-dependent enzymes could be classified as either acetyl- (or propionyl-) CoA carboxylases, pyruvate carboxylases, or oxaloacetate (or other dicarboxylate) decarboxylases (see table 1). Indeed, the experimentally characterized pyruvate carboxylases from *M. thermoautotrophicum* and *M. jannaschii* [Mukhopadhyay et al., 1998; Mukhopadhyay et al., 2000] appear to be the only biotin-dependent enzymes in these organisms, given that neither acetyl-CoA carboxyltransferase (AccD and AccA) subunits nor the β-subunit of oxaloacetate decarboxylase are encoded in their genomes (table 1). Likewise, the absence of *pycB* (*oadA*) and *oadB* genes in the genomes of *Halobacterium* sp. and *Sulfolobus solfataricus* leaves acetyl-CoA carboxylase as the only biotin-dependent enzyme in these archaea (table 1).

In several more complex cases, functional assignments for the biotin-dependent enzymes could be made on the basis of the analysis of the respective operons. Thus the aq_1520 gene that encodes a PycB-BCCP fusion protein from *Aquifex aeolicus* is located next to, and probably forms an operon with the biotin carboxylase-encoding gene aq_1517 (*accC*). The products of the aq_1517 and aq_1520 genes can be predicted to comprise a functional pyruvate carboxylase, similar to the one encoded by *M. jannaschii* genes MJ1229 and MJ1231 [Mukhopadhyay et al., 2000]. In contrast, no *accC*-like genes are located in the vicinity of PycB-BCCP fusion-encoding genes PH0834 of *Pyrococcus horikoshii* and aq_1614 of *A. aeolicus*, which makes their annotation more complex. In fact, the *A. aeolicus* genome does not encode an OadB subunit (see table 1), indicating that the aq_1614 protein cannot be a part of an oxaloacetate decarboxylase-like enzyme. It could potentially interact with aq_1470 or aq_1664 gene products, forming a pyruvate carboxylase, or with aq_445 and aq_1206 gene products, forming a methylmalonyl-CoA:pyruvate transcarboxylase, similar to the one found in *P. freudenreichii*, *G. intestinalis* and *E. histolytica*. In contrast, *P. horikoshii* does not encode the BC domain (table 1), indicating that PH0834 protein cannot be part of a pyruvate carboxylase. Given that both *P. horikoshii* and *Pyrococcus abyssi* contain operons (PH1283-PH1287 and PAB1769-PAB1772) that encode likely methylmalonyl-CoA decarboxylases (fig. 1; table 1), PH0834 would have to interact with some of their gene products to form a functional enzyme, either an oxaloacetate decarboxylase (with PH1283) or a methylmalonyl-CoA:pyruvate transcarboxylase (with PH1287). In the *Thermotoga maritima* genome, the *accB* gene is not fused to any other gene but forms a predicted operon with *accD/accA* gene. Since *T. maritima* does not encode a BC domain (table 1), this ACCT-BCCP complex could potentially interact with either the TM0880 gene product, forming a methylmalonyl-CoA decarboxylase, or with TM0128, forming a methylmalonyl-CoA:pyruvate transcarboxylase. Finally, the genome of *Archaeoglobus fulgidus* encodes the same components as *T. maritima* with the addition of the BC domain (AF0220). Various interactions of these components would allow this organism to assemble any of the known biotin-dependent enzymes. In such cases, genomic context data need to be supplemented with phylogenetic data that would help tracing the evolution of each particular domain and might explain its current function(s) in each particular organism (see below).

Evolutionary Relationships of Pyruvate Carboxylases and Oxaloacetate Decarboxylases

Since the carboxy-terminal PycB-BCCP part of the *Giardia* fusion protein shows similarity to both pyruvate carboxylases and oxaloacetate decarboxylases, we sought to investigate the evolutionary history of this gene fusion. In particular, it was important to determine whether the

Material may be protected by copyright law (Title 17, U.S. Code)

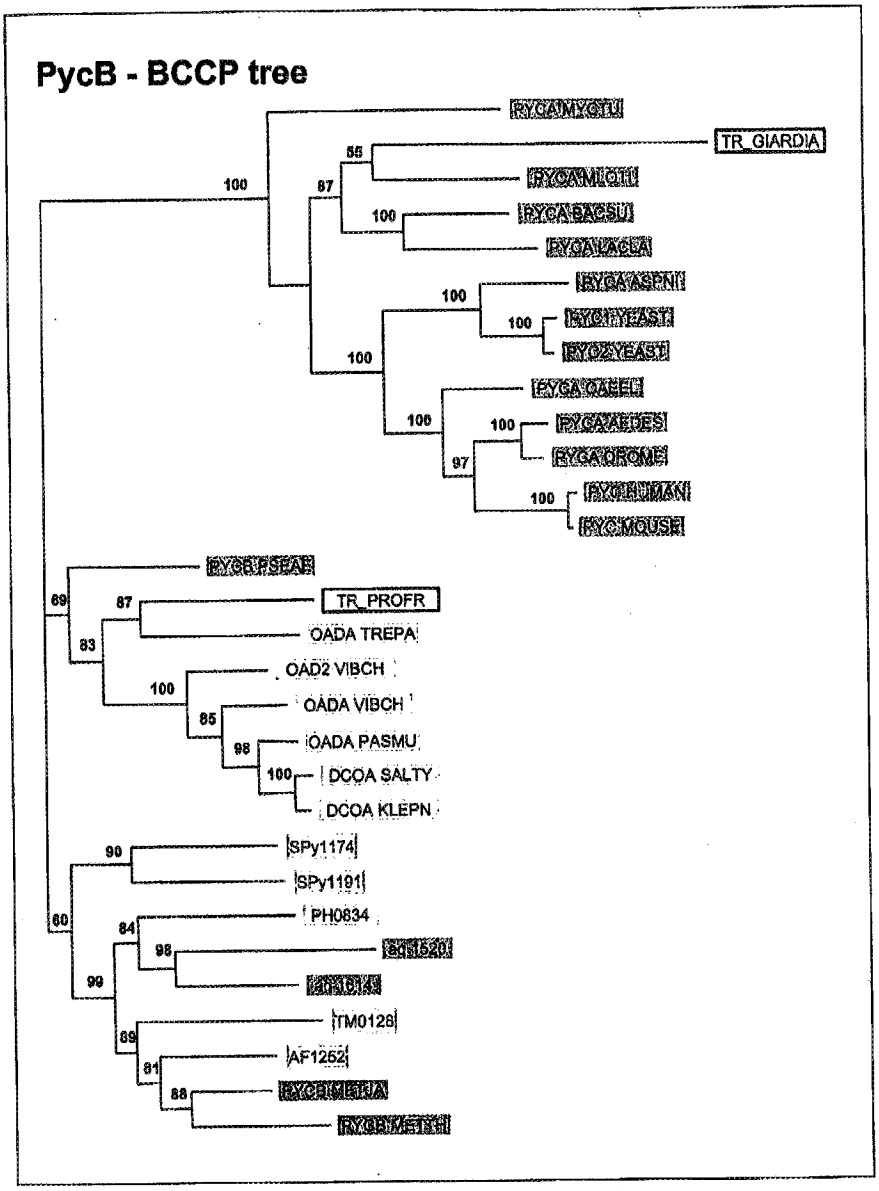
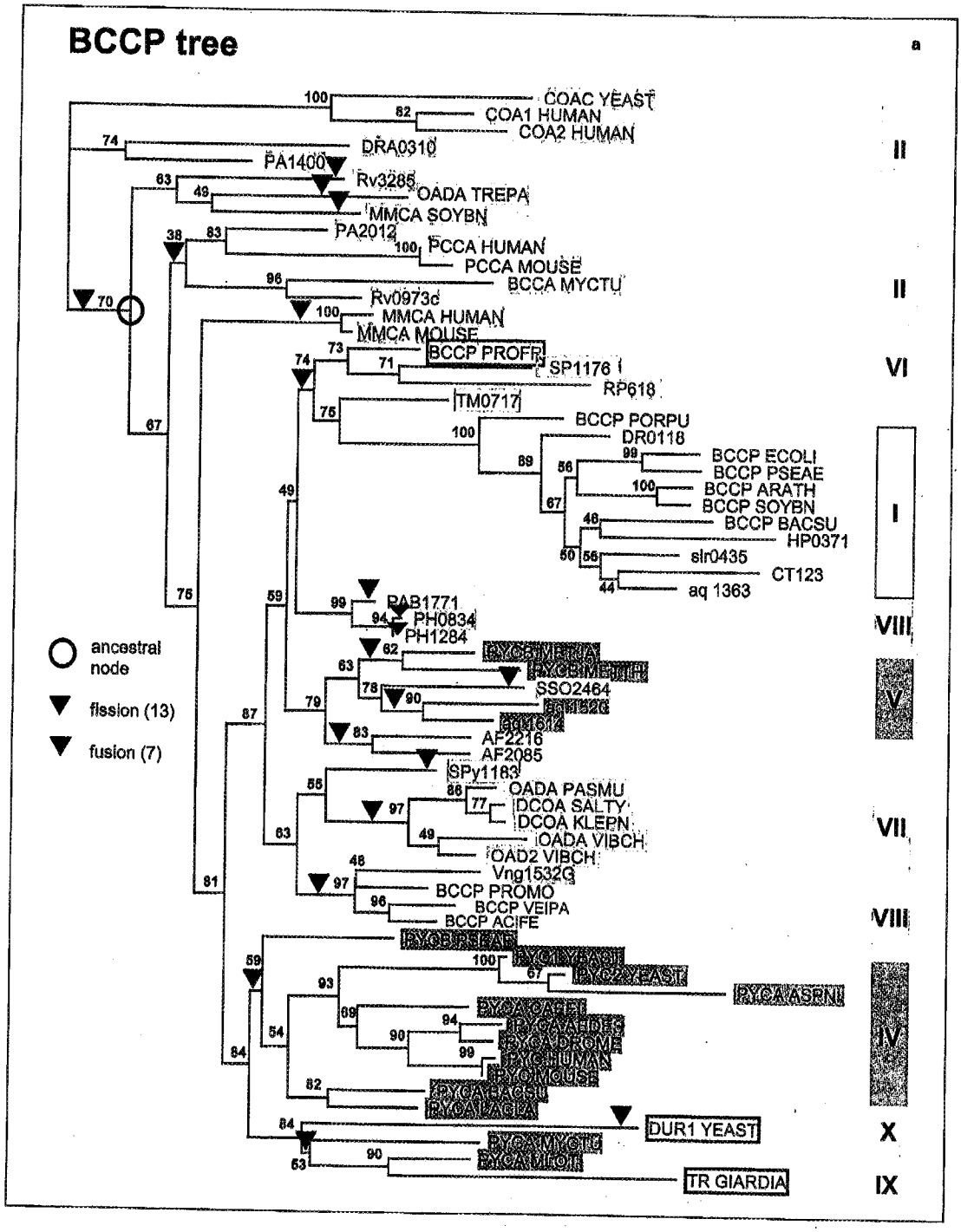


Fig. 2. A phylogenetic tree of the PycB-BCCP fusion proteins from various organisms. Tree reconstructions were performed as described in the Experimental Procedures section. The proteins are listed under their names in SWISS-PROT database, where available, or their names in complete genomes. SWISS-PROT-like names are as follows: PYCA_MYCTU, Rv2967c; PYCA_MLOTI, Mll3969; PYCA_LACLA, L63652; PYCA_ASPNI, CAC19838; PYCA_CAEEL, AAF60326; PYCA_AEDES, AAB64306; PYCA_DROME, AAD34740; PYCB_PSEAE, PA5435; OADA_TREPA, TP0056; OADA_VIBCH, VC0550; OAD2_VIBCH, VC0793m; OADA_PASMU, PM1422. Pyruvate carboxylases are shown in pink, oxaloacetate carboxylases are shown in yellow; transcarboxylases from *Giardia intestinalis* (TR_GIARDIA) and *Propionibacterium freudenreichii* (TR_PROFR) are framed in red. Unshaded proteins in yellow frames represent single PycB domains of predicted oxaloacetate decarboxylases that were, for the alignment purposes, artificially fused to the corresponding BCCP domains: TR_PROFR (5S and 1.3S subunits), Spy1174 (Spy1174-Spy1176), Spy1191 (Spy1191-Spy1183), TM0128 (TM0128-TM0717), and AF1252 (AF1252-AF2085). The numbers indicate percent bootstrap values (≥40%) for internal branches.

PycB-BCCP fusion in the *Giardia* biotin-dependent enzyme shared a common origin with similar fusions in pyruvate carboxylases and/or oxaloacetate decarboxylases. Alternatively, PycB and BCCP domains of the *Giardia* fusion protein could be most closely related to the corresponding subunits of the *P. freudenreichii* transcarboxylase, which has the same domain composition as the *Giardia* enzyme. To this end, we constructed phylogenetic trees of the PycB-BCCP fusions and of the individual BCCP and PycB domains from representatives of all major phylogenetic lineages (fig. 2, 3a, and 3b). For comparison, PycB-BCCP pairs from *P. freudenreichii* (5S and 1.3S subunits), *A. fulgidus* (AF1252-AF2085), *T. maritima* (TM0128-TM0717), and *Streptococcus pyogenes* (SPy1174-SPy1176 and Spy1191-Spy1183) were added to the alignment and shown on the PycB-BCCP tree (fig. 2).

The phylogenetic tree of the PycB-BCCP fusions (fig. 2) showed a tripartite division into: (1) bacterial and eukaryotic pyruvate carboxylases, (2) oxaloacetate decarboxylases, and (3) a group of biotin-dependent enzymes from thermophilic bacteria and archaea. The available experimental data [Mukhopadhyay et al., 1998; 2000] and functional inferences discussed above indicate that

Material may be protected by copyright law (Title 17, U.S. Code)



the third group included both archaeal *(M. thermoautotrophicum, M. jannaschii)* and bacterial *(A. aeolicus)* pyruvate carboxylases and archaeal *(P. horikoshii)* oxaloacetate decarboxylase or transcarboxylases. Remarkably, PycB-BCCP pairs from *A. fulgidus* (AF1252-AF2085) and *T. maritima* (TM0128-TM0717) also mapped into the third group, as did both pairs from *S. pyogenes* (fig. 2). Examination of the domain organization of the pyruvate carboxylases of the first and the third groups revealed a clear split in their domain organization. All bacterial and eukaryotic pyruvate carboxylases in the first group were three-domain fusion proteins with N-ter-

Material may be protected by copyright law (Title 17, U.S. Code)

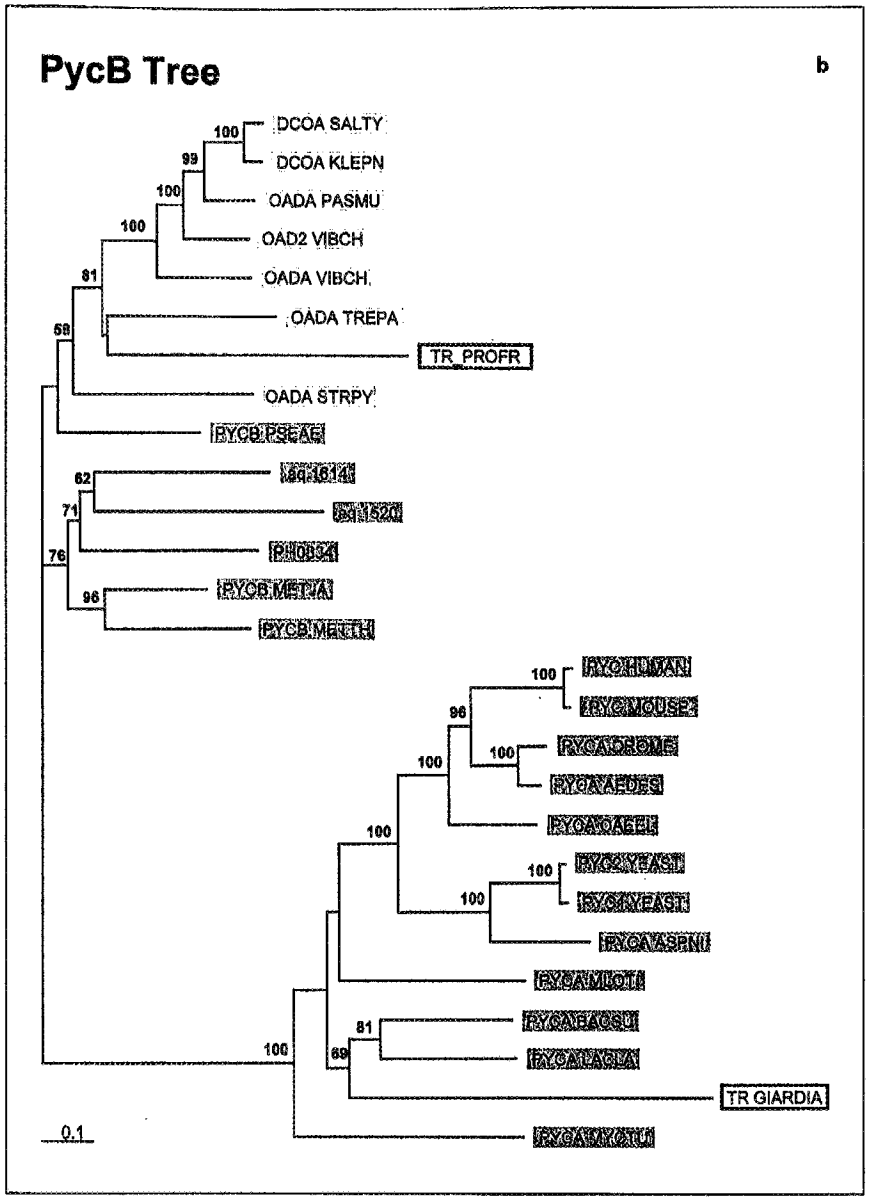


Fig. 3. Phylogenetic trees of the BCCP domain (**a**) and PycB domain (**b**). Protein names and coloring of pyruvate carboxylases and oxaloacetate decarboxylases is as in figure 2. The numbers indicate percent bootstrap values (≥40%) for internal branches. **a** BCCP domains of acetyl CoA carboxylase and other acyl-CoA carboxylases are colored blue. Red frames indicate BCCP domains of transcarboxylases (TR_GIARDIA, BCCP_PROFR) and urea amidolyase (DUR1_YEAST). Stand-alone BCCP subunits are unshaded. BCCP subunits of methylmalonyl-CoA decarboxylases from *Propionigenium modestum* and *Veillonella parvula* and of glutaconyl-CoA decarboxylase from *Acidaminococcus fermentans* are shown as BCCP_PROMO, BCCP_VEIPA, and BCCP_ACIFE, respectively. The bars and Roman numerals on the right indicate domain architectures as shown in figure 1. The ancestral node is indicated by the circle. Red triangles indicate putative domain fission events; blue triangles indicate putative domain fusion events.

minal biotin carboxylase domains (domain structure IV in fig. 1). In contrast, PycB-BCCP fusions in the third group did not contain any extra domains, although, in some cases, their genes were located next to the genes encoding the biotin carboxylase subunit. Remarkably, the *Giardia* PycB-BCCP fusion protein clustered with the first group of pyruvate carboxylases, suggesting that this enzyme could have evolved from a BC-PycB-BCCP fusion, but definitely did not originate from a fusion of *P. freudenreichii* subunits (fig. 2).

Evolution of the PycB-BCCP Enzymes: One or Multiple Fusions?

Because of the complex structure of the PycB-BCCP phylogenetic tree we were interested in finding out whether, during evolution of biotin-dependent enzymes, fusion of these two domains could have occurred more than once. This possibility was first investigated by comparing the topologies of phylogenetic trees reconstructed from separate alignments of the BCCP and PycB domains (fig. 3a, b). In particular, we sought to determine whether pyruvate carboxylase and oxaloacetate decarboxylase

Material may be protected by copyright law (Title 17, U.S. Code)

evolved via a single, ancestral PycB-BCCP fusion event, followed by dissemination of the fusion gene by vertical descent and/or horizontal gene transfer, or via multiple, independent fusion events. The BCCP domain phylogeny is potentially most interesting with respect to this question because it reflects the evolutionary relationships among BCCP sequences that are present in several distinct genetic contexts. In this tree, the BCCP domains found in PycB-BCCP fusions do not appear to be monophyletic, which suggests the possibility of multiple, perhaps three or four, independent PycB-BCCP domain fusions that led to the formation of pyruvate carboxylases and oxaloacetate decarboxylases, respectively (fig. 3a). However, BCCP is a small domain, which results in a small number of phylogenetically informative sites in the multiple alignment and, consequently, low bootstrap support for many nodes in the tree (fig. 3a). Therefore the conclusion of multiple independent PycB-BCCP fusions is only tenuously supported by the phylogenetic reconstruction for BCCP.

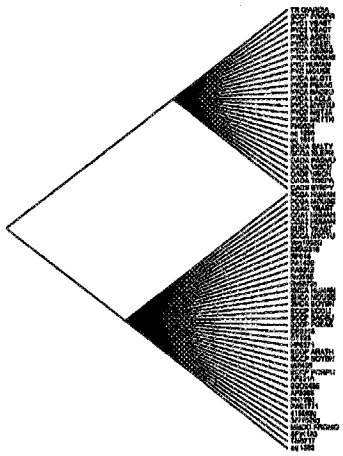
In such cases, phylogenetic hypothesis testing can be employed to further assess evidence for or against specific evolutionary scenarios [Hillis et al., 1996; Jordan et al., 2001; Xia, 2000]. This approach involves generation of constraint trees that represent specific phylogenetic hypotheses. In particular, a constraint tree was generated that forced all of the BCCP domains encoded by genes with PycB-BCCP fusions to be monophyletic (fig. 4). The BCCP alignment was then used to reconstruct the best phylogeny that was consistent with this constraint tree. The log likelihood score of the resulting best (constrained) hypothetical tree was compared to the score of the best overall (unconstrained) maximum likelihood tree using the Kishino-Hasegawa test. The results of this comparison support the hypothesis that the BCCP domains from PycB-BCCP fusions are not monophyletic because the best constraint tree with enforced monophyly of BCCP domains from these fusions was significantly less likely than the best maximum likelihood tree where BCCP domains from PycB-BCCP fusions are polyphyletic (fig. 4). The hypothesis testing procedure was repeated four times, with each successive iteration using a constraint tree progressively more similar to the best BCCP tree. In other words, each successive constraint tree conformed less stringently to the requirement of monophyly of BCCP domains from PycB-BCCP fusions. For example, the second constraint tree (fig. 4) allowed these BCCP domains to form two monophyletic groups, one for the pyruvate carboxylases and one for the oxaloacetate decarboxylases; such a tree would be consistent with two independent PycB-BCCP fusions. Trees 3–5 in figure 4 correspond to different phylogenies with 3 or more independent fusions. Despite this progressive relaxation of constraints, each of which resulted in a phylogeny consistent with a greater number of independent PycB-BCCP fusions than the preceding one, all these hypothetical BCCP phylogenies were significantly worse than the best BCCP tree (fig. 4). Thus, analysis of constrained trees appears to support at least 5 independent PycB-BCCP fusion events leading to the emergence of pyruvate carboxylases and oxaloacetate decarboxylases.

An alternative to multiple, independent fusions as the evolutionary scenario for PycB-BCCP enzymes might involve multiple fission events. By mapping potential domain fusion and/or fission events on a phylogenetic tree, it is possible to derive the most parsimonious scenario (one with the minimal number of steps), assuming that different types of evolutionary events, in this case fusions and fissions, are equally likely. Given the BCCP domain phylogeny, the ancestral state for the two domains in question can be hypothetically assigned as either fused or not fused. Once so assigned, the topology of the tree can be used to determine the number of domain fusions or fissions required to produce the observed distribution of BCCP domains fused to PycB and those that are partners in other fusions or stand-alone proteins. If the ancestral state was assumed to be a PycB-BCCP fusion, 13 subsequent fission events would be required to yield the observed phyletic distribution of BCCP domains (fig. 3a). If the ancestral state was represented by separate PycB and BCCP domains, the BCCP phylogeny suggests 7 independent fusion events (fig. 3a). Thus, under the assumption of equal likelihoods of fusion and fission events, the scenario of multiple, independent fusions is the more parsimonious one. Since gene fusion appears to be even somewhat more common in evolution than gene fission [Snel et al., 2000], it appears most likely that independent PycB-BCCP fusions indeed took place in the course of evolution of biotin-dependent enzymes. The above analy-

Fig. 4. Phylogenetic hypothesis testing comparing the best overall BCCP tree with BCCP constraint trees that represent specific evolutionary scenarios. Five different constraint trees, each of which represents a specific phylogenetic hypothesis (see Results) regarding PycB-BCCP domain fusions, are shown. These unresolved phylogenies were used to constrain maximum likelihood tree reconstructions such that the best tree (the one with the highest log likelihood score) that is consistent with the constraint tree will be generated. The log likelihood scores of the best constraint trees are then statistically compared to the log likelihoods of the best overall (unconstrained) tree using the Kishino-Hasegawa test.

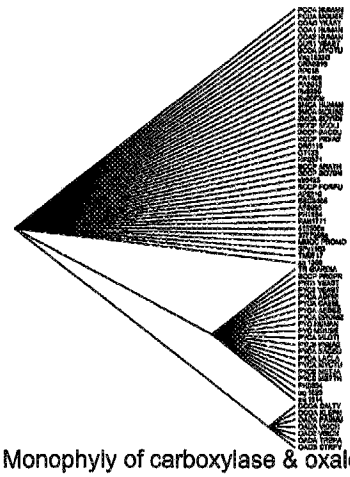
Material may be protected by copyright law (Title 17, U.S. Code)

1



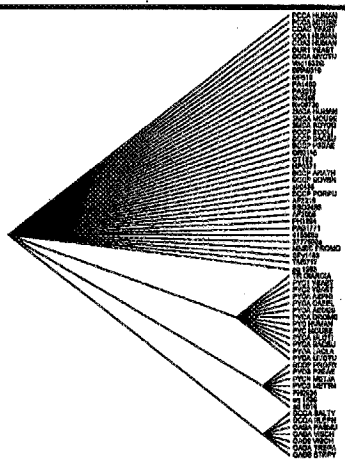
Monophyly of all PycB-BCCP fusions

2



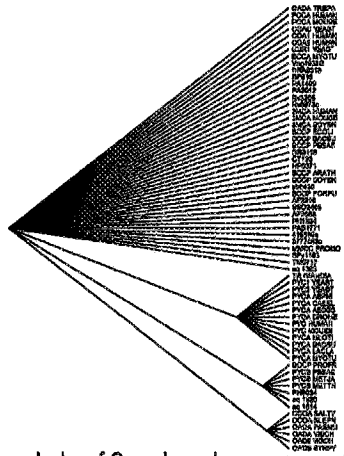
Monophyly of carboxylase & oxaloacetate PycB-BCCP fusions

3



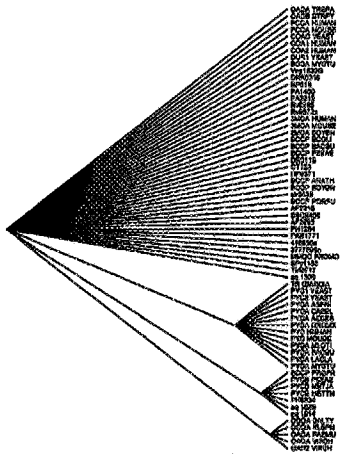
Monophyly of 2 carboxylase groups & oxaloacetate PycB-BCCP fusions

4



Monophyly of 2 carboxylase groups & oxaloacetate (except TREPA) PycB-BCCP fusions

5



Monophyly of 2 carboxylase groups & oxaloacetate (except TREPA & STRPY) PycB-BCCP fusions

Phylogenetic hypothesis testing

best BCCP tree versus constraint trees 1–5
(each constraint tree is less restrictive than the previous one)

Statistical comparison of trees

Hasegawa & Kishino [1989]

Tree	log L	diff ± (SE)	p
best	-6960.83		
1	-7043.04	82.21 ± 20.85	8.0×10^{-5}
2	-7039.57	78.74 ± 21.99	3.4×10^{-4}
3	-7038.78	77.95 ± 19.47	6.2×10^{-5}
4	-7022.44	61.61 ± 18.47	8.5×10^{-4}
5	-7017.54	56.71 ± 16.80	7.3×10^{-4}

Material may be protected by copyright law (Title 17, U.S. Code)

Table 2. PycB phylogenetic comparisons

Tree	log L[1]	Differenz ± SE[2]	p[3]
PycB[4]	−16860.36		
Monophyletic[5]	−16865.60	5.25 ± 6.73	0.4 (n.s.)
BCCP[6]	−16968.79	110.17 ± 23.83	3.8×10^{-6}

1 Log likelihood scores for each of the three trees.
2 Difference between the log likelihood scores for the PycB tree and either the monophyletic or the BCCP topologies with the standard error of the difference.
3 Probability that the difference between the log likelihoods of the topologies being compared is due to chance.
4 Best PycB tree topology (unconstrained).
5 Constrained tree topology with monophyly of carboxylases and decarboxylases.
6 BCCP tree topology.

sis is over-simplified in that the actual history of these enzymes probably involved both fusions and fissions. Another potential problem with this analysis is the low bootstrap values associated with many of the internal nodes in the BCCP tree (fig. 3a). A more conservative approach involved collapsing the unsupported nodes and then counting the number of fusion or fission events (depending upon the hypothetical ancestral state) in the resulting partially unresolved tree. Under this approach, the multiple independent fusion scenario remains the more parsimonious one (6 fusions versus 9 fissions; data not shown).

The topology of the PycB phylogenetic tree (fig. 3b) was different from that of the BCCP tree (fig. 3a), based on the Kishino-Hasegawa test (table 2). If the two domains had fused once in a common ancestor and then evolved as a linked unit (via speciation and/or horizontal transfer), then the two domain phylogenies would be expected to have the same topology. The incongruence of the two domain trees seems to lend additional support to the hypothesis of independent PycB-BCCP fusion events.

The co-evolutionary dynamics of the PycB and BCCP domains fused in various biotin-dependent enzymes seems to reflect a trend of multiple independent fusions between pairs of domains giving rise to similar enzymes. A similar pattern has been detected in a recent evolutionary analysis of the fusion between Cu-chaperones and Cu-ATPases [Jordan et al., 2001]. The repeated generation and maintenance of similar domain architectures of enzymes through independent fusions probably reflects selective pressure that favors tightly regulated co-expression of the fused domains to facilitate their assembly and substrate channeling between these domains as compared to the assembly of different subunits. Thus, independent domain fusions represent cases of convergent evolution at the level of protein domain architecture.

Evolution of the BCCP-Biotin Carboxylase Fusions

A similar problem of discriminating between a single ancestral fusion and multiple independent fusion events was addressed for the BC-BCCP domain fusions that are found in both acetyl-CoA carboxylases and propionyl-CoA carboxylases. We also sought to determine whether the current diversity of BC-BCCP fusions is consistent with this fusion occurring just once in the evolutionary history of various biotin-dependent carboxylases. The BCCP tree, with its abundance of domains from different genic contexts, is suited to addressing this question as well. If there was a single ancestral BC-BCCP fusion, then the BCCP domains found in these fusions should be monophyletic in the BCCP tree. The observation that this is not the case (fig. 3a) argues for multiple independent BC-BCCP domain fusions. Because the low bootstrap values of the BCCP tree render this conclusion tenuous, as described above, phylogenetic hypothesis testing was performed to more rigorously assess the evidence for each of the two alternative scenarios. In the case of the BC-BCCP fusion, whole-tree statistical comparisons did not yield any significant difference between the best BCCP tree and the hypothetical trees that are consistent with a single ancestral fusion event (data not shown). Thus, in this case, the phylogenetic evidence is not robust enough to clearly discriminate between the two evolutionary scenarios for the evolution of the BC-BCCP fusion and the more parsimonious scenario of a single fusion event appears to be plausible.

Using Phylogeny to Improve Functional Assignment

Because, as shown above, the individual domains of biotin-dependent enzymes are prone to shuffling, fusion and fission, their functions are bound to change, too, depending on the particular genome context. This makes analysis of gene neighborhoods (operon structure and domain fusions) a powerful way to deduce the function of each particular protein. Phylogenetic analysis of individual domains could also help in improving the functional assignments by providing a time coordinate and therefore allowing one to identify relatively recent fusions that might have unusual functions.

Consider, for example, the biotin-dependent enzymes of the syphilis spirochete *Treponema pallidum*. Although,

Material may be protected by copyright law (Title 17, U.S. Code)

in this organism, the PycB-BCCP fusion (TP0056, see table 1) is encoded next to the OadB and OadG subunits and definitely functions as a subunit of an oxaloacetate decarboxylase [Häse et al., 2001], *T. pallidum* also encodes a BC subunit (TP0695), but no ACCT subunit (table 1). It follows that the BC subunit can only interact with the PycB-BCCP fusion to form a pyruvate carboxylase. Although, in phylogenetic trees, the PycB domain of *T. pallidum* TP0056 definitely clusters with the PycB domains of other oxaloacetate decarboxylases (fig. 3b), its BCCP domain belongs to the cluster of typical bacterial stand-alone BCCP domains (fig. 3a). This unusual phylogenetic position of the *T. pallidum* BCCP domain indicates that the PycB-BCCP fusion in *T. pallidum* is a relatively recent event, which most likely occurred through recombination of PycB- and BCCP-encoding genes, followed by elimination of the original BCCP domain (or subunit) of the oxaloacetate decarboxylase. Similar multiple recombination events have been recently observed in the evolution of the lipoyl-binding domain, which is distantly related to the BCCP [Omelchenko et al., 2002]. It appears that the ancestral form of this spirochete encoded both a biotin-dependent carboxylase of acetyl-CoA or pyruvate and an oxaloacetate decarboxylase. Genome constriction during the evolution of this parasite led to the formation of the novel PycB-BCCP domain fusion, TP0056, which apparently functions in both reactions, decarboxylation of oxaloacetate and carboxylation of pyruvate. The annotation of the TP0056 protein probably should reflect both those possibilities.

Another case where an apparently recent PycB-BCCP domain fusion has led to a change of function is the *P. horikoshii* protein PH0834, the only instance of the PycB domain in the three species of pyrococci sequenced so far (see table 1). The PycB domain of PH0834 clusters with the PycB domains of pyruvate carboxylases from *Aquifex* (fig. 3b), whereas its BCCP domain is closely related to the two stand-alone BCCP subunits of the pyrococcal methylmalonyl-CoA decarboxylase (fig. 3a). This suggests that PH0834, too, is a relatively recent PycB-BCCP fusion that emerged in *P. horikoshii* after acquisition of the *pycB* gene through lateral gene transfer from a different lineage of thermophiles. This new protein provided *P. horikoshii* with a new functionality – oxaloacetate decarboxylation – that is missing in its relatives, *Pyrococcus furiosus* and *P. abyssi* (see table 1).

Finally, both PycB and BCCP domains of the *Giardia* enzyme clearly fall into the pyruvate carboxylase group (fig. 3a, b). This suggests that the C-terminal portion of this protein has evolved from a pyruvate carboxylase ancestor, most likely a protomitochondrial enzyme, similar to the pyruvate carboxylase of the α-proteobacterium *Mesorhizobium loti* (see fig. 3a). Transfer of the corresponding *pycB* gene into the nucleus, followed (or preceded) by its recombination with the gene(s) encoding ACCT domains gave rise to this unique domain combination. The phylogenetic data clearly show that this domain organization cannot be explained by a simple fusion of the three subunits of the *P. freudenreichii* transcarboxylase. The presence of ACCT-PycB fusions in distantly related protozoa, *Entamoeba* and *Giardia*, suggests that this fusion has either occurred very early in the evolution of eukaryotes or was spread by lateral gene transfer.

Discussion

Biotin-dependent enzymes catalyze several key reactions in cellular metabolism [Buckel, 2001; Wakil et al., 1983; Wallace et al., 1998]. Reactions catalyzed by pyruvate carboxylase (EC 6.4.1.1) and acetyl-CoA carboxylase (EC 6.4.1.2) comprise the first steps, respectively, in gluconeogenesis and fatty acid biosynthesis, and propionyl-CoA carboxylation (EC 6.4.1.3) is a key step in the metabolism of odd-numbered long-chain fatty acids. Hydrolysis of urea to ammonia and carbon dioxide, catalyzed by urea amidolyase (EC 6.3.4.6), also begins with biotin-dependent carboxylation of urea [Roon et al., 1972]. 3-Methylcrotonyl-CoA carboxylase participates in the leucine catabolism pathway in plant mitochondria [Anderson et al., 1998]. The reverse reaction, biotin-dependent decarboxylation of oxaloacetate (EC 4.1.1.3), methylmalonyl-CoA (EC 4.1.1.41) and glutaconyl-CoA (EC 4.1.1.70) is coupled with the generation of transmembrane sodium gradient in many anaerobic bacteria and comprises the principal mechanism of energy transformation in these organisms [Buckel, 2001; Dimroth and Schink, 1998]. In addition, glutaconyl-CoA decarboxylation is an important step in microbial degradation of benzoyl-CoA and other aromatic compounds [Müller and Schink, 2000]. Finally, a transcarboxylase (EC 2.1.3.1) from *P. freudenreichii* catalyzes transfer of the carboxyl moiety from methylmalonyl-CoA to pyruvate [Wood and Zwolinski, 1976; Wood and Kumar, 1985].

The domain organization and evolution of biotin-dependent enzymes have been the subject of several previous studies that have uncovered the complexity of the domain order and fusion patterns in this enzyme family [Li and Cronan, 1992; Samols et al., 1988; Toh et al., 1993; Tu and Hagedorn, 1997]. Studies of the *P. freuden-*

Material may be protected by copyright law (Title 17, U.S. Code)

reichii methylmalonyl-CoA:pyruvate transcarboxylase by the Wood group defined the three subunits of this enzyme, namely, acetyl-CoA carboxyltransferase (12S) subunit, pyruvate carboxyltransferase (5S) subunit, and the BCCP (1.3S) subunit [Wood, 1979; Wood and Kumar, 1985]. They have also recognized BCCP as the common subunit (or domain) in all biotin-dependent carboxylases and PycB as the common subunit (or domain) in pyruvate carboxylases and oxaloacetate decarboxylases and suggested that the biotin-containing enzymes have evolved from a common ancestor that has adapted to a variety of metabolic environments [Samols et al., 1988]. Li and Cronan [1992] cloned and sequenced *Escherichia coli accD* and *accA* genes and demonstrated the similarity of their products to the N-terminal and C-terminal parts, respectively, of rat propionyl-CoA carboxylase, thus bridging the prokaryotic and eukaryotic carboxylases. Subsequent analysis of the evolution of biotin carboxylases and BCCPs, initiated by Toh et al. [1993] and developed later by Tu and Hagedorn [1997], came up with identical phylogenetic trees for both these domains (or subunits). This observation indicated that BCs and BCCPs have evolved as a single unit (co-evolved) in eukaryotic acetyl-CoA carboxylases, propionyl-CoA carboxylases, and pyruvate carboxylases. In all these enzymes, co-evolution of BCs and BCCPs could be attributed to the fusion of the corresponding genes to produce a single polypeptide chain (see fig. 1). In contrast, evolution of BCs and BCCPs in bacterial acetyl-CoA carboxylases and pyruvate carboxylases, where the corresponding genes are not necessarily fused into a single gene (fig. 1), remained largely unexplored. In addition, the recently developed methods for protein domain analysis, which led to the creation of protein domain databases, such as Pfam [Bateman et al., 2002], CDD [Marchler-Bauer et al., 2002], and COGs [Tatusov et al., 2000], were not available to investigators previously. This led, for example, to Toh and coworkers [Toh et al., 1993] questioning homology between the AccA-like domains of eukaryotic acetyl-CoA and propionyl-CoA carboxylases, which has been noticed even earlier [Li and Cronan, 1992], but became compelling with the development of novel sequence analysis methods. Furthermore, the abundance of new sequence data from complete genome sequences justified a re-analysis of the evolution of biotin-dependent enzymes.

Because of the multi-domain organization of biotin-dependent enzymes and the wide range of the reactions catalyzed by them, straightforward sequence comparisons do not always allow accurate functional predictions for proteins encoded by newly sequenced genes. Instead or in addition, functional assignments for such genes must also draw from the subunit organization of the deduced enzyme or, in the case of prokaryotes, from the gene arrangement in the corresponding operons. When genes encoding different subunits of a biotin-dependent enzyme do not form a single operon (which is the case, for example, for the *P. freudenreichii* transcarboxylase and for the *E. coli* acetyl-CoA carboxylase), this approach may also become problematic [e.g. Mukhopadhyay et al., 2000]. Therefore, in the absence of experimental data for the newly sequenced ORFs in various genomes, it would be reasonable to refrain from assigning putative enzymatic activities. Instead, functions of individual domains and/or domain combinations can be assigned with reasonable confidence.

In contrast, when domains are fused into a single multi-domain protein, assignment of the enzyme activity becomes relatively straightforward. Nevertheless, understanding the metabolic role(s) of a particular enzyme, such as, for example, the role of methylmalonyl-CoA:pyruvate transcarboxylase in the central metabolism of *G. intestinalis* and *E. histolytica*, remains a challenge. These two organisms belong to separate eukaryotic lineages, and the evolutionary origin of their peculiar metabolic make-up is not yet resolved. Recent studies indicate that both lineages descend from mitochondrion-containing ancestors via secondary structural and functional losses [Roger, 1999]. In addition, both organisms harbor metabolic enzymes that are generally not found in other eukaryotes [Field et al., 2000; Sanchez et al., 2000]. In *P. freudenreichii*, the likely function of methylmalonyl-CoA:pyruvate transcarboxylase involves balancing gluconeogenesis with fatty acid biosynthesis. However, at this time there is no evidence that gluconeogenesis exists in either *Giardia* or *Entamoeba*, and *Giardia* is apparently incapable of fatty acid biosynthesis; specifically, acetate from acetyl-CoA is not incorporated into *Giardia* cell lipids [Jarroll et al., 1989]. Fatty acid degradation has not been described in *Giardia* either. Indeed, a search of the *Giardia* genome database [McArthur et al., 2000] shows the presence of the *fadA* but not the *fadB* gene, which encode α and β subunits, respectively, of the fatty acid oxidation complex. An interesting alternative would be that the ACCT-PycB enzyme catalyzes conversion of pyruvate into oxaloacetate, whose decarboxylation back to pyruvate by a membrane-bound oxaloacetate decarboxylase would then be coupled with the generation of sodium-motive force on the cytoplasmic membrane, as described for many anaerobic bacteria [Buckel, 2001; Dimroth and Schink, 1998; Häse et al., 2001]. No OadB homologs could be found in

Material may be protected by copyright law (Title 17, U.S. Code)

the current version of the *Giardia* genome database [McArthur et al., 2000], although *Entamoeba* appears to encode one (data not shown). However, the *Giardia* genome sequence is far from being complete [McArthur et al., 2000], so there is still a chance that *Giardia* encodes the *fadB* or *oadB* gene (or both). In conclusion, the ACCT-PycB-BCCP fusion enzyme can be expected to play a role in the central metabolism of *G. intestinalis* and *E. histolytica*, but its specific function remains to be established. It should be noted that, in contrast to *G. intestinalis* and *E. histolytica, Leishmania major* encodes a typical eukaryotic pyruvate carboxylase [Myler et al., 1999].

As an early-diverging eukaryote, *Giardia* could serve as an interesting object for understanding the evolutionary forces underlying domain fusion in eukaryotes. The unusual biotin-dependent enzyme studied here and the recently described acetyl-CoA synthetase (ADP-forming) are just two examples of unusual multidomain proteins in *Giardia*. It can be expected that further progress of the *G. intestinalis* genome-sequencing project [McArthur et al., 2000] reveals many more unexpected domain architectures.

Experimental Procedures

Cloning and Sequencing of the Giardia intestinalis *Transcarboxylase*

In the course of a general search of the *Giardia intestinalis* genome database at the Josephine Bay Paul Centre, Woods Hole, Ma. [McArthur et al., 2000], for enzymes involved in central carbohydrate metabolism and its accessory pathways, the keyword 'pyruvate' identified a genomic clone (MJ3007RA) that contained a sequence (annotated as pyruvate kinase) with high similarity to bacterial pyruvate carboxylases. For instance, the expectation value of the BLASTP [Altschul et al., 1997] hit occurring by chance (E-value) with the *Bacillus stearothermophilus* enzyme was 6×10^{-68}. Since pyruvate carboxylase activity in *G. intestinalis* had never been described and its role in *Giardia* metabolism was elusive, we decided to clone and sequence this gene for further analysis.

Two primers, Glpc1 (5′-GAATGCCCACCAATCATTGC-3′) and Glpc2 (5′-GTAACCACTGCATTTGCG-3′), corresponding to amino acid residues 653–659 and 733–739 of the final sequence, respectively, were used to amplify a 258-bp fragment from genomic DNA of *G. intestinalis* in a standard PCR reaction. This fragment was inserted into the TA cloning vector (Invitrogen, Carlsbad, Calif., USA) and its identity was verified by sequencing. The fragment was randomly labeled to specific activity of about 10^7 cpm/μg. One million recombinants of the genomic *G. intestinalis* library were screened with this probe and two positive clones, gGlpc3 and gGlpc5, were identified. Both clones were sequenced on both strands by primer walking. The gGlpc5 clone contained the 5′ part of the final sequence, whereas the gGlpc3 clone contained its 3′ part, with both clones overlapping in the region between nucleotides 1300 and 2412 (corresponding to the amino acid residues 433–804) of the complete coding sequence. The complete open reading frame was 4,017 bp long and encoded a 1,338-aa protein. It was deposited in GenBank (NCBI, NIH, Bethesda, Md., USA) with the accession number AF207740.

Sequence Analysis

The multiple alignment of the biotin carboxyl carrier proteins and BCCP domains from multidomain proteins was constructed on the basis of the Pfam [Bateman et al., 2002] alignment for biotin-lipoyl domain, PF00364, and expanded according to the results of PSI-BLAST searches [Altschul et al., 1997] and the three-dimensional structure of BCCP [Athappilly and Hendrickson, 1995; Yao et al., 1997]. Multiple alignments of the PycB, AccD, AccA, and BC domains were constructed from the ClustalW alignments from the COG database [Tatusov et al., 2000] by manual adjustment on the basis of the results of PSI-BLAST searches [Altschul et al., 1997]. The alignments used for phylogenetic analyses are available at the NCBI ftp site (ftp://ftp.ncbi.nih.gov/pub/galperin/BiotinAln.html).

Phylogenetic Analysis

Phylogenetic trees were constructed using neighbor-joining and maximum likelihood (ML) methods. The presented results are based on the ML analyses. Evolutionary distances between aligned protein sequences were calculated with the Dayhoff PAM matrix as implemented in the PROTDIST program of the PHYLIP package [Felsenstein, 1996]. Neighbor-joining phylogenies [Saitou and Nei, 1987] were constructed from the resulting distances using the NEIGHBOR program of PHYLIP. ML phylogenies were constructed using the ProtML program of the MOLPHY package [Adachi and Hasegawa, 1992]. The JTTF model of amino acid substitutions [Jones et al., 1992] was used with local re-arrangements in the ML analysis to optimize neighbor-joining seed trees. For each ML phylogeny, 10,000 bootstrap replicates were performed by using the resampling of estimated log-likelihoods method implemented in ProtML [Adachi and Hasegawa, 1992; Kishino et al., 1990]. Constraint trees for distance-based phylogenetic hypothesis testing were generated using the MacClade v3.01 [Maddison and Maddison, 1992] program. Constraint trees were imported into PAUP* [Swofford, 1998] and used to constrain the reconstruction of neighbor-joining trees, built with user-supplied PAM protein distances (determined as described above), to generate the phylogenies for alternative hypotheses. The resulting phylogenies were compared to the best overall trees via ML using the Kishino-Hasegawa test [Kishino and Hasegawa, 1989] implemented in the TREEPUZZLE program [Strimmer and von Haeseler, 1996]. Alternative topologies used for ML-based phylogenetic hypothesis testing were generated manually by modifying the original tree using the Treeview program [Page, 1996]. Alternative trees were compared to the best tree using the Kishino-Hasegawa test implemented in ProtML.

Acknowledgments

Sequencing of the *G. intestinalis* transcarboxylase gene was performed at the laboratory of Professor Miklós Müller (Laboratory of Biochemical Parasitology, The Rockefeller University, New York) and was supported by NIH grant AI11942 to Miklós Müller and an EMBO fellowship to Katrin Henze. We thank Miklós Müller for support and helpful comments.

Material may be protected by copyright law (Title 17, U.S. Code)

References

Adachi, J. and Hasegawa, M. 1992. MOLPHY: Programs for Molecular Phylogenetics. Tokyo, Institute of Statistical Mathematics.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zheng, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST – A new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Anderson, M.D., Che, P., Song, J., Nikolau, B.J., and Wurtele, E.S. 1998. 3-Methylcrotonyl-coenzyme A carboxylase is a component of the mitochondrial leucine catabolic pathway in plants. Plant Physiol. 118:1127–1138.

Athappilly, F.K. and Hendrickson, W.A. 1995. Structure of the biotinyl domain of acetyl-coenzyme A carboxylase determined by MAD phasing. Structure 3:1407–1419.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. Nucleic Acids Res. 30:276–280.

Biagini, G.A., Lloyd, D., Kirk, K., and Edwards, M.R. 2000. The membrane potential of *Giardia intestinalis*. FEMS Microbiol Lett. 192: 153–157.

Bouzat, J.L., McNeil, L.K., Robertson, H.M., Soltcr, L.F., Nixon, J.E., Beever, J.E., Gaskins, H.R., Olsen, G., Subramaniam, S., Sogin, M.L., and Lewin, H. A. 2000. Phylogenomic analysis of the alpha proteasome gene family from early-diverging eukaryotes. J. Mol. Evol. 51:532–543.

Brown, D.M., Upcroft, J.A., Edwards, M.R., and Upcroft, P. 1998. Anaerobic bacterial metabolism in the ancient eukaryote *Giardia duodenalis*. Int. J. Parasitol. 28:149–164.

Bruderer, T., Wehrli, C., and Kohler, P. 1996. Cloning and characterization of the gene encoding pyruvate phosphate dikinase from *Giardia duodenalis*. Mol. Biochem. Parasitol. 77: 225–233.

Buckel, W. 2001. Sodium ion-translocating decarboxylases. Biochim. Biophys. Acta 1505:15–27.

Cronan, J.E. Jr. and Waldrop, G.L. 2002. Multi-subunit acetyl-CoA carboxylases. Prog. Lipid Res. 41:407–435.

Dimroth, P. 1997. Primary sodium ion translocating enzymes. Biochim. Biophys. Acta 1318:11–51.

Dimroth, P. and Schink, B. 1998. Energy conservation in the decarboxylation of dicarboxylic acids by fermenting bacteria. Arch. Microbiol. 170:69–77.

Embley, T.M. and Hirt, R.P. 1998. Early branching eukaryotes? Curr. Opin. Genet. Dev. 8:624–629.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266: 418–427.

Field, J., Rosenthal, B., and Samuelson, J. 2000. Early lateral transfer of genes encoding malic enzyme, acetyl-CoA synthetase and alcohol dehydrogenases from anaerobic prokaryotes to *Entamoeba histolytica*. Mol. Microbiol. 38: 446–455.

Galperin, M.Y. and Koonin, E.V. 1998. Sources of systematic error in functional annotation of genomes: Domain rearrangement, non-orthologous gene displacement, and operon disruption. In Silico Biol. 1:55–67.

Häse, C.C., Fedorova, N.D., Galperin, M.Y., and Dibrov, P.A. 2001. Sodium ion cycle in bacterial pathogens: Evidence from cross-genome comparisons. Microbiol. Mol. Biol. Rev. 65: 353–370.

Hayashi, M., Nakayama, Y., and Unemoto, T. 2001. Recent progress in the Na^+-translocating NADH-quinone reductase from the marine *Vibrio alginolyticus*. Biochim. Biophys. Acta 1505:37–44.

Henze, K., Morrison, H.G., Sogin, M.L., and Müller, M. 1998. Sequence and phylogenetic position of a class II aldolase gene in the amitochondriate protist, *Giardia lamblia*. Gene 222: 163–168.

Hilario, E. and Gogarten, J.P. 1998. The prokaryote-to-eukaryote transition reflected in the evolution of the V/F/A-ATPase catalytic and proteolipid subunits. J. Mol. Evol. 46:703–715.

Hillis, D.M., Mable, B.K., and Moritz, C. 1996. Applications of molecular systematics: The state of the field and a look to the future. In Hillis, D.M., Moritz, C., and Mable, B.K. (eds): Molecular Systematics. Sunderland, Sinauer Associates, pp 515–543.

Jarroll, E.L., Manning, P., Berrada, A., Hare, D., and Lindmark, D.G. 1989. Biochemistry and metabolism of *Giardia*. J. Protozool. 36:190–197.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–82.

Jordan, I.K., Natale, D.A., Koonin, E.V., and Galperin, M.Y. 2001. Independent evolution of heavy-metal-associated domains in copper chaperones and copper-transporting ATPases. J. Mol. Evol. 53:622–633.

Kishino, H. and Hasegawa, M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29:170–179.

Kishino, H., Miyata, T., and Hasegawa, M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151–160.

Kondo, H., Kazuta, Y., Saito, A., and Fuji, K. 1997. Cloning and nucleotide sequence of *Bacillus stearothermophilus* pyruvate carboxylase. Gene 191:47–50.

Kulda, J. and Nohýnková, E. 1995. Giardia in humans and other animals. In Kreier, J.P. (ed): Parasitic Protozoa, ed 2. San Diego, Academic Press, vol 10, pp 225–422.

Li, S.J. and Cronan, J.E. Jr. 1992. The genes encoding the two carboxyltransferase subunits of *Escherichia coli* acetyl-CoA carboxylase. J. Biol. Chem. 267:16841–16847.

Lindmark, D.G. 1980. Energy metabolism of the anaerobic protozoon *Giardia lamblia*. Mol. Biochem. Parasitol. 1:1–12.

Maddison, D.R. and Maddison, W.P. 1992. MacClade. Sunderland, Sinauer Associates.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., and Bryant, S.H. 2002. CDD: A database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 30: 281–283.

Martin, W. and Müller, M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392: 37–41.

McArthur, A.G., Morrison, H.G., Nixon, J.E., Passamaneck, N.Q., Kim, U., Hinkle, G., Crocker, M.K., Holder, M.E., Farr, R., Reich, C.I., et al. 2000. The *Giardia* genome project database. FEMS Microbiol. Lett. 189:271–273.

Morris, C.P., Lim, F., and Wallace, J.C. 1987. Yeast pyruvate carboxylase: Gene isolation. Biochem. Biophys. Res. Commun. 145:390–396.

Morrison, H.G., Roger, A.J., Nystul, T.G., Gillin, F.D., and Sogin, M.L. 2001. *Giardia lamblia* expresses a proteobacterial-like DnaK homolog. Mol. Biol. Evol. 18:530–541.

Mukhopadhyay, B., Stoddard, S.F., and Wolfe, R.S. 1998. Purification, regulation, and molecular and biochemical characterization of pyruvate carboxylase from *Methanobacterium thermoautotrophicum* strain deltaH. J. Biol. Chem. 273:5155–5166.

Mukhopadhyay, B., Patel, V.J., and Wolfe, R.S. 2000. A stable archaeal pyruvate carboxylase from the hyperthermophile *Methanococcus jannaschii*. Arch. Microbiol. 174:406–414.

Mukhopadhyay, B. and Purwantini, E. 2000. Pyruvate carboxylase from *Mycobacterium smegmatis*: Stabilization, rapid purification, molecular and biochemical characterization and regulation of the cellular level. Biochim. Biophys. Acta 1475:191–206.

Müller, J.A. and Schink, B. 2000. Initial steps in the fermentation of 3-hydroxybenzoate by *Sporotomaculum hydroxybenzoicum*. Arch. Microbiol. 173:288–295.

Müller, M. 1998. Enzymes and compartmentation of anaerobic protists – A special case in eukaryotic evolution? In Coombs, G.H., Vickerman, K., Sleigh, M.A., and Warren, A. (eds): Evolutionary relationships among protozoa. Dordrecht, Kluwer, pp 109–131.

Myler, P.J., Audleman, L., deVos, T., Hixson, G., Kiser, P., Lemley, C., Magness, C., Rickel, E., Sisk, E., Sunkin, S., Swartzell, S., Westlake, T., Bastien, P., Fu, G., Ivens, A., and Stuart, K. 1999. *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proc. Natl. Acad. Sci. USA 96:2902–2906.

Nevalainen, L., Hrdy, I., and Müller, M. 1996. Sequence of a *Giardia lamblia* gene coding for the glycolytic enzyme, pyruvate, phosphate dikinase. Mol. Biochem. Parasitol. 77:217–223.

Material may be protected by copyright law (Title 17, U.S. Code)

Omelchenko, M.V., Makarova, K.M., and Koonin, E.V. 2002. Recurrent intragenomic recombination leading to sequence homogenization during the evolution of the lipoyl-binding domain. FEMS Microbiol. Lett. 209:255–260.

Page, R.D.M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comp. Appl. Biosci. 12:357–358.

Park, J.H., Schofield, P.J., and Edwards, M.R. 1997. Pyruvate kinase is present in *Giardia intestinalis*. Exp. Parasitol. 87:153–156.

Reeves, R.E. 1984. Metabolism of *Entamoeba histolytica* Schaudinn, 1903. Adv. Parasitol. 23: 105–142.

Roger, A.J. 1999. Reconstructing early events in eukaryotic evolution. Am. Natur. 154 (suppl 4): S146–S163.

Roon, R.J., Hampshire, J., and Levenberg, B. 1972. Urea amidolyase. The involvement of biotin in urea cleavage. J. Biol. Chem. 247:7539–7545.

Rosenthal, B., Mai, Z., Caplivski, D., Ghosh, S., de la Vega, H., Graf, T., and Samuelson, J. 1997. Evidence for the bacterial origin of genes encoding fermentation enzymes of the amitochondriate protozoan parasite *Entamoeba histolytica*. J. Bacteriol. 179:3736–3745.

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

Samols, D., Thornton, C.G., Murtif, V.L., Kumar, G.K., Haase, F.C., and Wood, H.G. 1988. Evolutionary conservation among biotin enzymes. J. Biol. Chem. 263:6461–6464.

Sanchez, L.B., Morrison, H.G., Sogin, M.L., and Müller, M. 1999. Cloning and sequencing of an acetyl-CoA synthetase (ADP-forming) gene from the amitochondriate protist, *Giardia lamblia*. Gene 233:225–231.

Sanchez, L.B., Galperin, M.Y., and Müller, M. 2000. Acetyl-CoA synthetase from the amitochondriate eukaryote *Giardia lamblia* belongs to the newly recognized superfamily of acyl-CoA synthetases (nucleoside diphosphate-forming). J. Biol. Chem. 275:5794–5803.

Snel, B., Bork, P., and Huynen, M. 2000. Genome evolution. Gene fusion versus gene fission. Trends Genet. 200:9–11.

Sogin, M.L., Gunderson, J.H., Elwood, H.J., Alonso, R.A., and Peattie, D.A. 1989. Phylogenetic meaning of the kingdom concept: An unusual ribosomal RNA from *Giardia lamblia*. Science 243:75–77.

Strimmer, K. and von Haeseler, A. 1996. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964–969.

Suguri, S., Henze, K., Sanchez, L.B., Moore, D.V., and Müller, M. 2001. Archaebacterial relationships of the phosphoenolpyruvate carboxykinase gene reveal mosaicism of *Giardia intestinalis* core metabolism. J. Eukaryot. Microbiol. 48:493–497.

Sumper, M. and Riepertinger, C. 1972. Structural relationship of biotin-containing enzymes. Acetyl-CoA carboxylase and pyruvate carboxylase from yeast. Eur. J. Biochem. 29:237–248.

Swofford, D.L. 1998. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sunderland, Sinauer Associates.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33–36.

Toh, H., Kondo, H., and Tanabe, T. 1993. Molecular evolution of biotin-dependent carboxylases. Eur. J. Biochem. 215:687–696.

Triana, O., Galanti, N., Olea, N., Hellman, U., Wernstedt, C., Lujan, H., Medina, C., and Toro, G.C. 2001. Chromatin and histones from *Giardia lamblia*: A new puzzle in primitive eukaryotes. J. Cell. Biochem. 82:573–582.

Tu, Z. and Hagedorn, H.H. 1997. Biochemical, molecular, and phylogenetic analysis of pyruvate carboxylase in the yellow fever mosquito, *Aedes aegypti*. Insect Biochem. Mol. Biol. 27: 133–147.

Wakil, S.J., Stoops, J.K., and Joshi, V.C. 1983. Fatty acid synthesis and its regulation. Annu. Rev. Biochem. 52:537–579.

Wallace, J.C., Jitrapakdee, S., and Chapman-Smith, A. 1998. Pyruvate carboxylase. Int. J. Biochem. Cell Biol. 30:1–5.

Wood, H.G. and Zwolinski, G.K. 1976. Transcarboxylase: Role of biotin, metals, and subunits in the reaction and its quaternary structure. CRC Crit. Rev. Biochem. 4:47–122.

Wood, H.G. 1979. The anatomy of transcarboxylase and the role of its subunits. CRC Crit. Rev. Biochem. 7:143–160.

Wood, H.G. and Kumar, G.K. 1985. Transcarboxylase: Its quaternary structure and the role of the biotinyl subunit in the assembly of the enzyme and in catalysis. Ann. N.Y. Acad. Sci. 447:1–22.

Wu, G., McArthur, A.G., Fiser, A., Sali, A., Sogin, M.L., and Müller, M. 2000. Core histones of the amitochondriate protist, *Giardia lamblia*. Mol. Biol. Evol. 17:1156–63.

Xia, X. 2000. Data analysis in molecular biology and evolution. Boston, Kluwer.

Yao, X., Wei, D., Soden, C., Summers, M.F., and Beckett, D. 1997. Structure of the carboxy-terminal fragment of the apo-biotin carboxyl carrier subunit of *Escherichia coli* acetyl-CoA carboxylase. Biochemistry 36:15089–15100.

Material may be protected by copyright law (Title 17, U.S. Code)