Evolutionary Analysis of γ-Carbonic Anhydrase and Structurally Related Proteins

Gustavo Parisi, María Fornasari, and Julián Echave

Centro de Estudios e Investigaciones, Universidad Nacional de Quilmes, Roque Sáenz Peña 180, 1876 Bernal, Argentina

Received July 6, 1998; revised May 28, 1999

We studied the evolutionary relationships between γ -carbonic anhydrase (γ -CA) and a very diverse group of proteins that share the sequence motif characteristic of the left-handed parallel β -helix (L β H) fold. This sequence motif is characterized by the imperfect tandem repetition of short hexapeptide units, which makes it difficult to obtain a reliable alignment based on sequence information alone. To solve this problem, we used a structural alignment of three members of the group with known crystallographic structures as a seed to obtain a reliable sequence alignment. Then, we applied protein maximum-parsimony and maximumlikelihood phylogenetic inference methods to this alignment. We found that y-CA belongs to a diverse superfamily of proteins that share the LBH domain. This superfamily is composed mainly of acyltransferases. The most remarkable feature of the phylogenetic tree obtained is that its main branches group together functionally related proteins, so that the coarse topology can be rather easily explained in terms of functional diversification. Regarding the main branch of the tree containing γ -CA, we found that, in addition to the group of its closest relatives that had already been studied, y-CA is closely related to the tetrahydrodipicolinate N-succinyltransferases. © 2000 Academic Press

INTRODUCTION

The enzyme γ -carbonic anhydrase (γ -CA) catalyzes the reversible hydration of carbon dioxide. It belongs to a functionally convergent group of carbonic anhydrases, composed of three unrelated members (α -, β -, and γ -carbonic anhydrases), present in animals, plants, archaea, and eubacteria. It has been shown that the γ -CA from the methanogenic archaeon *Methanosarcina thermophila* (Alber and Ferry, 1994) exhibits no significant sequence similarity to α -CA and β -CA (Hewett-Emmett and Tashian, 1996).

The functional form of γ -CA is a homotrimer with three zinc-containing active sites located at the interfaces between two monomers. The monomer fold, shown in Fig. 1a, is formed by a left-handed parallel β -helix (L β H) topped by a short α -helix (Kisker *et al.*, 1996).

The L_βH fold exhibits a characteristic sequence pattern that folds as a left-handed spiral around the surface of an equilateral triangular prism. The $L\beta H$ sequence pattern is composed of the imperfect tandem repetition of short hexapeptide units. Each helical wound is composed of three hexapeptide units. Each hexapeptide unit is a sequence motif termed hexapeptide repeat (Vuorio et al., 1994) or isoleucine patch (Dicker and Seetharam, 1992). The annotation of the hexapeptide repeat in the PROSITE database (Bairoch, 1993) is [LIV]-[GAED]-X₂-[STAV]-X, where X stands for any amino acid. The hexapeptide repeat is characterized by an aliphatic residue, usually Ile, Val, or Leu at the first position, that we shall call *i* in what follows. A small residue (Ala, Ser, Cys, Val, Thr, or Asn) is found at position i + 4. Another well-conserved residue is glycine at positions i + 1.

The L β H sequence pattern has been found in a large and diverse group of acyltransferases. Members of this group are involved in a variety of enzymatic processes, such as amino acid metabolism, cell wall biosynthesis in microorganisms, and antibiotic neutralization (Anderson and Raetz, 1987; Vuorio *et al.*, 1991; Downie, 1989; Murray and Shaw, 1997). Although it is recognized that γ -CA has borderline similarities to some members of this group of L β H-containing enzymes (Hewett-Emmett and Tashian, 1996), the relationship of γ -CA with this group has not been studied in detail.

In this paper we examine the evolutionary relationships of γ -CA with the L β H acyltransferases. There are two obstacles to overcome in performing such a study. First, it is difficult to align sequences characterized by very short sequence repetitions. Second, the genes studied are so divergent that DNA-based phylogenetic inference is unreliable due to saturation effects. To surmount these obstacles, we apply phylogenetic inference methods to a protein sequence alignment based on a structural alignment of the known structures of some members of the group studied.

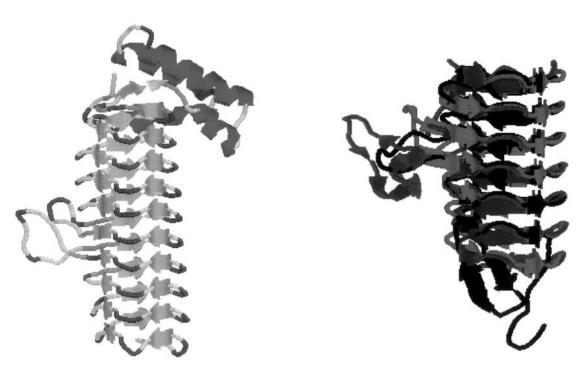
MATERIALS AND METHODS

Similarity Searches

Preliminary sequence similarity searches were performed using BLAST (Altschul et al., 1990), FASTA3







d

FIG. 1. Plots of the left-handed parallel β -helix (L β H) domain of the three proteins of Table 1 used for the structural alignment. (a) 1THJ, γ -carbonic anhydrase from *Methanosarcina thermophila* (Kisker *et al.*, 1996); (b) 1TDT, tetrahydrodipicolinate *N*-succinyltransferase from *Mycobacterium bovis* (Beaman *et al.*, 1997); (c) 1LXA, UDP-*N*-acetylglucosamine acyltransferase from *Escherichia coli* (Raetz and Roderick, 1995); and (d) multiple structural alignment of the three L β H domains, obtained with STAMP. These plots were obtained using RASMOL (Roger and Milner-White, 1995).

(Pearson and Lipman, 1988), and the motif-searching method PROBE (Neuwald et al., 1997). The resulting high-score sequences were used to build a multiple alignment using CLUSTAL W version 1.6 (Thompson et al., 1994). With this alignment we built profiles (Gribskov et al., 1987) with PROFILEMAKE (Lüthy et al., 1994), using different substitution matrices and weighting schemes (logarithmic and proportional to the distance). These profiles were used to search the Swiss Protein database (Swiss-Prot) release 35 (Bairoch and Boeckmann, 1992) with PROFILESEARCH using Zscore = 7 as a cutoff. The new high-score proteins were incorporated into the initial alignment and the procedure was iterated until no new sequences were obtained. Similarity percentages were calculated using the pairwise alignment option in the AMPS program (Barton, 1990) and using PAM250, with gap penalty = 10 and with 100 randomizations of each sequence pair, to evaluate the reliability of each alignment.

Sequence Alignment

From the sequences recovered from the similarity searches, we selected those with known crystallographic structures. Then, we performed a *structural alignment* with STAMP version 4.0 (Russell and Barton, 1992) using PAIRWISE and TREEWISE with the SCAN option invoked.

We incorporated the rest of the sequences recovered from the similarity searches into the structural alignment using the profile option in CLUSTAL W. The alignment was adjusted manually using GENDOC (Nicholas and Nicholas, 1997) to maximize the conserved positions. Since the structural alignment was used as a seed for the alignment process, the alignment obtained will be called *structure-based alignment* to differentiate it from the structural alignment and from a standard alignment based on sequence information alone. In this study, only the left-handed parallel β -helix domains of each sequence were analyzed. Therefore, the amino and carboxy termini were deleted to fit the length to the structural alignment.

Phylogenetic Analysis

Saturation analysis. The structure-based alignment was converted into a DNA alignment with PUT-GAPS (McInerney, 1997) and the corresponding DNA sequences downloaded from the GenBank database release 106 (Benson *et al.*, 1998). Saturation was studied graphically by plotting, for each sequence pair, the percentage of transitions and transversions against sequence divergence. Sequence divergence was estimated using Kimura's two-parameter distance (Kimura, 1980) and the maximum-likelihood method (Felsenstein, 1981), both of them employing the DNADIST program from the PHYLIP package Version 3.57c (Felsenstein, 1993).

Phylogenetic inference. Maximum-parsimony and protein maximum-likelihood methods were used to infer the phylogenetic relationships. To perform parsimony analysis we used the PROTPARS program from PHYLIP. The large size of our data set made an exhaustive search impossible. Therefore, parsimony trees were obtained using the JUMBLE option of PHYLIP, with 10 replicates. Bootstrap resampling (1000 times) was carried out to quantify the relative support of the branches of the inferred trees. A majority-rule consensus tree was obtained using the CONSENSE program included in PHYLIP.

The topology of the branch containing γ -CA was further explored by performing a maximum-likelihood exhaustive search. We used PROTML, from MOLPHY 2.2 (Adachi and Hasegawa, 1994), with the JTT transition probability matrix (Jones *et al.*, 1992) corrected for the amino acid frequencies observed in the data set studied. Estimated bootstrap confidence was calculated by the resampling estimated log-likelihood (RELL) method (Hasegawa and Kishino, 1994). The support at each node was obtained using mol2con (perl script provided by Arlin Stolzfus). We also calculated the relative-likelihood support (RLS) using Treecons (Jermin *et al.*, 1997) with a class V weighting scheme and an α value of 0.01.

RESULTS

Proteins can be referred to by their complete name, Swiss-Prot ID codes, gene name, or activity. To make the discussion as clear as possible, we found it convenient to use a context-dependent designation. See Table 1 for further reference.

Similarity Searches

Starting with the sequence of *M. thermophila* γ -CA (CAH_METTE), the similarity searches recovered only its closest relatives. These are FBP PSEAE (see Gen-Bank Accession No. M82832 and Swiss-Prot Accession No. P40882) and Y304 METJA, improperly identified as ferripyochelin-binding proteins; CAIE_ECOLI, belonging to the carnitine operon of *Escherichia coli*; CCMM_SYNP7, a CO_2 -concentrating mechanism protein; and YRDA ECOLI, a hypothetical protein of E. coli. The relationship of most of these proteins with γ -CA has been already reported by other authors (Alber and Ferry, 1994; Hewett-Emmett and Tashian, 1996). The previous set of sequences was used to build a sequence profile to start the iterative similarity search already described. From these searches we recovered the 58 sequences shown in Table 1. We note for further reference that the minimal similarity between pairs of sequences was 21%.

Sequence Alignment

Three proteins included in this set have known three-dimensional structure: γ -CA from *M. thermo*-

TABLE 1

$L\beta H$ Superfamily

Activity	PDB ID	Swiss-Protein ID	Gene	Source	Branch
UDP-N-acetylglucosamine acyltransferase	1LXA	LPXA_BRUAB LPXA_CHRVI LPXA_ECOLI LPXA_HAEIN LPXA_PROMI LPXA_RICRI LPXA_SALTY_	LpxA LpxA LpxA LpxA LpxA LpxA LpxA	Brucella abortus Chromatium vinosum Escherichia coli Haemophilus influenzae Proteus mirabilis Rickettsia rickettsii Salmonella typhimurium	A1 A1 A1 A1 A1 A1 A1 A1
UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acyl- transferase		LPXA_YEREN LPXD_ECOLI LPXD_HAEIN LPXD_RICRI LPXD_SALTY LPXD_YEREN	LpxA LpxD LpxD LpxD LpxD LpxD LpxD	Yersinia enterocolitica Escherichia coli Haemophilus influenzae Rickettsia rickettsi Salmonella typhimurium Yersinia enterocolitica	A1 A2 A2 A2 A2 A2 A2
Chloramphenicol acetyltransferase		CAT4_AGRTU CAT4_ECOLI CAT4_ENTAE CAT4_MORMO	Cat Cat CatB4 Cat	Agrobacterium tumefaciens Escherichia coli Enterobacter aerogenes Morganella morganii	B1 B1 B1 B1
Streptogramin A-acetyltransferase Virginiamycin A-acetyltransferase Probable macrolide acetyltransferase		SATA_ENTFC VATA_STAAU MATA_BACSH	SatA Vat ErmG	Enterococcus faecium Staphylococcus aureus Bacillus sphaericus	B2 B2 B2 B2
Bifunctional: UDP-N-acetylglucosamine pyrophos- phorylase and Glucosamine-1-phosphate acetyltrans- ferase UDP-N-acetylglucosamine pyrophosphorylase		GLMU_ECOLI GLMU_HAEIN GLMU_NEIGO GCAD_BACSU	GlmU GlmU GlmU GcaD	Escherichia coli Haemophilus influenzae Neisseria gonorrhoeae Bacillus subtilis	C C C C
Galactoside-O-acetyltransferase		THGA_ECOLI	LacA	Escherichia coli	D
Galactoside O-acetyltransferase (Putative)		THGA_LACLA WBBJ_ECOLI	LacA WbbJ	Lactococcus lactis Escherichia coli	D D
Maltose O-acetyltransferase		YJV8_YEAST MAA_ECOLI	Yj1218w MaA	Saccharomyces cerevisiae Escherichia coli	D D
Nod factors O-acetyl transferase		YYAL_BACSU (*) NODL_RHILV	YyaL NodL NodL	Bacillus subtilis Rhizobium leguminosarum Rhizobium malilati	D D D
Acetyltransferase (Putative) Involved in biosynthesis of slime polysacharid colanic acid		NODL_RHIME WCAB_ECOLI WCAF_ECOLI	NodL WcaB WcaF	Rhizobium meliloti Escherichia coli Escherichia coli	D D D
Acetyltransferase (Putative)		YA39_SCHPO	Spac18b11.09c	$Schizos accharomyces\ pombe$	D
Serine acetyltransferase		CYSE_BACSU CYSE_BUCAP CYSE_ECOLI CYSE_HAEIN CYSE_HELPY CYSE_SALTY CYSE_STAXY CYSE_SYNP7 CYSE_SYNP7 CYSE_SYNP3 SRPH_SYNP7 NIFP_AZOCH	CysE CysE CysE CysE CysE CysE CysE CysE	Bacillus subtilis Buchnera aphidicola Escherichia coli Haemophilus influenzae Helicobacter pylori Salmonella typhimurium Staphylococcus xylosus Synechococcus sp. (strain pcc 7942) Synechococcus sp. (strain pcc 6803) Synechococcus sp. (strain pcc 7942) Azotobacter chroococcum mcd 1	Н Н Н Н Н Н Н Н Н Н Н Н Н Н Н Н Н Н Н
Carbonic anhydrase CO2-concentrating mechanism protein Unknown Improperly called Ferrypiochelin binding protein Unknown Unknown Involved in Carnitine biosynthesis	1THJ	CAH_METTE CCMM_SYNP7 Y304_METJA(*) FBP_PSEAE	Cam CcmM Mj0304 Fbp	Methanosarcina thermophila Synechococcus sp. (strain pcc 7942) Methanococcus jannaschii Pseudomonas aeruginosa	F1 F1 F1 F1
		YRDA_ECOLI (*) CAIE_ECOLI	YrdA CaiE	Escherichia coli Escherichia coli	F1 $F1$
Unknown Involved in the biosynthesis of TAB toxine		TABB_PSESZ	TabB	Pseudomonas syringae	F2
Involved in the biosynthesis of TAB toxine Tetrahydrodipicolinate N-succinyltransferase	1TDT	DAPD_ACTPL DAPD_ECOLI DAPD_HAEIN DAPD_MYCBO	DapD DapD DapD DapD	Actinobacillus pleuropneumoniae Escherichia coli Haemophilus influenzae Mycobacterium bovis	F2 F2 F2 F2
Unknown Involved in biosynthesis of type 1 capsular polysaccha- rides		CAPG_STAAU	CapG	Staphylococcus aureus	—

Note. List of the 58 proteins recovered from the similarity searches. In the first column we show the activity found in the databases Swiss-Prot, KEGG (Kanehisa, 1996), and/or WIT (Selkov *et al.*, 1996). In the last column we show the main branch and/or subbranch of the phylogenetic tree (Figs. 5 and 6) to which each protein belongs. * Hypothetical proteins.

phila (Fig. 1a; Kisker et al., 1996), tetrahydrodipicolinate N-succinyltransferase from Mycobacterium bovis (Fig. 1b; Beaman et al., 1997), and UDP-N-acetylglucosamine acyltransferase from E. coli (Fig. 1c; Raetz and Roderick, 1995). The STAMP structural alignment, shown in Fig. 1d, produced an alignment of 176 residues with a structural similarity score (Sc) of 5.44 and a final root mean square deviation of 0.85. The structural alignment included the entire left-handed β -helix domain of γ -CA and tetrahydrodipicolinate N-succinyltransferase (7 complete helical coils for γ -CA, 5 complete and 2 partial helical coils for tetrahydrodipicolinate N-succinyltransferase) and almost the whole UDP-N-acetylglucosamine acyltransferase left-handed β -helix domain (9 of 10 helical coils). In Fig. 2 we show the sequence alignment corresponding to the structural alignment of Fig. 1d. It is noteworthy that the main gaps observed correspond to loops inserted at the corners of the coils, such as those in positions 69 to 83 and 99 to 108 in UDP-N-acetylglucosamine acyltransferase, 166 to 175 and 210 to 224 in tetrahydrodipicolinate N-succinyltransferase, and 60 to 64 and 81 to 105 in γ -CA.

Attempts to perform a multiple alignment of the 58 sequences with CLUSTAL W, using sequence information alone, produced results that depend strongly on the order of the sequences in the input. Furthermore, in all cases the alignment of the three sequences with known structures, described in the previous paragraph, are wrong compared with the more reliable structural alignment of Fig. 2. Therefore, to obtain a more reliable alignment, we used the structural alignment of Fig. 2 as a seed to obtain the structure-based sequence alignment shown in Fig. 3.

Phylogenetic Analysis

The graphical DNA saturation analysis of Fig. 4 clearly shows saturation in transitions and transversions. For this reason, phylogenetic inference procedures were applied directly to the protein alignment of Fig. 3. As a result of the maximum-parsimony analysis, we found two equally parsimonious trees. In Fig. 5 we show the corresponding consensus tree. We should note that the two most-parsimonious trees differ only in the topology of the serine acyltransferase family (Fig. 5, branch E).

The most remarkable feature of the phylogenetic tree is that functionally related proteins are clustered together, forming the main branches. This can be easily seen in Table 1, where the proteins studied are grouped by function and sorted by branch. Although a detailed analysis of each main branch is beyond the scope of the

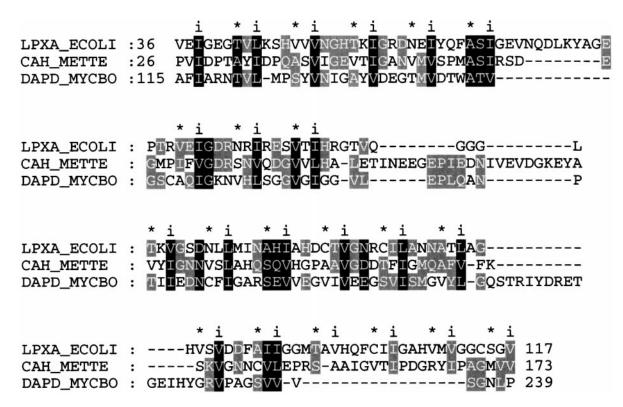


FIG. 2. Sequence alignment resulting from the structural alignment of Fig. 1d. Columns *i* and i + 4 of the hexapeptide repeat units (see text) are indicated explicitly with an i and *, respectively. Numbers at the beginning and end of each sequence indicate the position of each sequence in the complete protein. Conserved positions are shaded using GENDOC with 90, 75, and 50% conserved, using PAM250 as scoring table. The Swiss-Protein ID is used to identify each sequence. See Table 1 for further reference.

666	DRGAL- DRGAL- GG- GG- GG- CG- CG- CG- CG- CG- CG- CG	KGDTVIG EGNVTIG EGNVTIG EGNVEG EGNVEG KGVQIG STTPIPVHHEL STTPIPVHEL STTAHPIDVA DPMSSPNIPPD OPMSSPNIPPD	TTABHPDDFEQ	
INTOSFLAYAH AHDCD SHNYMAFSHAGHDCV SHNYMAFSHAGHDCV NNNL MINNH AHDCT NNNL MINNH AHDCT NNNL MINNH AHDCT NNNL MINAH AHDCT SNL MINAH AHDCT SNL MINAH AHDCT R-VI GNNVE GANTC I -VK GNNVE GANTC	CLU CORVECTATION CONTRACT S-VH CORVECTACT S-VH CORVECTACT PEFENNSYLPACTAC EPFENSYLDAFORACTA EPFESALDAFORACTA EPFESALDAFORACTA EPFESALDAFORACTA	THUNDER CHARLES	DETAL FRANK TYTAHPDI NR	GERVEN ANGOOT W GRUVT W GRUV-ALL CHN-ALL CHN-ALL GHC-A
- ARG YTS TRC TRC 		м- НТП - ГЕН - ГЕН - ГЕР - ГЕ		- VV -LETINEGEPIEDNIVEVDGKETAVV VGR
VNCITREGVT#HKG - SDN	CAUGE FAN- CAUGE FAN- GADGE GYAN CADGE GYAN CEPEF FYMCE SEPEF YMCE SEPEF YMCE FYFPHLFRMG	-ANIHRMDC TTYPFNIFGCG -ANIHRMDC TTYPFNIFGCG WILRDPARFDIRGT -VTLRIPPARFDIRGT -VTLIDPMNTY SPD -VTLIDPMNTY SPD -VTLIDPMNTY SPD -VTLIDPMNTY UNDY STESVT IGROTLIAS - DNFYANNILVLIDGA DNFYANNILVLIDGA DNFYANNULVLIDGA DNFFVNFOCVLLDVC AMVYINNOVLDVC	IXKNVAGDFT RHG	
GCDPQNNKHKGG-PTR_NV_VNK GSEPQDLSFTPEKARP_I I ODH GEVNQDLKYAGE-PTR_EI ORN GEVNQDLKYAGE-PTR_EI ORN GEVNQDLKYAGE-PTQ_I I ONN GOPPOILKYAGE-PTR_EI ONN GOPPOILKYAGE-PTR_EI ONN GENNQDLKYAGE-PTR_PI ONN GANVTIYHE OI ONN MANVTIYHE OI ONN WANVTIYHE OI ONN		- VKF MM - - KL LLACV - KL LLACV - KL LLACV - KL - KL - - KL - - KN - 		- 10 HPG - 10 HPG - V0 HPG - V0 HPG - V0 H 10 - V1 10 - V0 110 - CSSN010 - CSSN010 - CSSN010 - CSSN010 - CSSN010 - CSSN010
TTGATTLGAGTKY YPHA II UFGVTRUGAHNR CHGATT WNGHTKLGRDNE YQF ASI UNGHTKLGRDNE YQF ASI UNGHTKLGRDNE YQF ASI UNGHTRLGRDNO YQF ASI TBGTTELGENTU YPF ASI UNGQTK IGCDNE YQF ASI UNGQTK IGCDNE YQF ASI UNGTKLGCRDNE YQF ASI UNGTKLGCRDNE YQF ASI UNGGTK IGCRDNE YQF ASI UNGGTY IGCRDNE YQF	LLEAGSFI /IGAGCFU /ILPD-EG /LLPDRDD /LFPRDD /LFPRDD /LYPIL	HYEF LOCKLIT BEGVNNRLQLSRLE JEGVNNRLQLSRLE JEGVNNRLQLAELE AGONNRRLQLAELE TGVDRVALSQAE KENFATVG - ENAMY KENFETG - ENAMY KENFETG - ENAMY KENESELG - ENAMY RELLGSVG - OCVTI ADLFGSVG - OCVTI EELLGSVG - OCVTI	SAEQWHLFLREGEVGETGEVG FOAVI RPFFLTHIO PLUVLYRNRAFLLRLEGAKIGKNUV RPFFTTHIO VLYRNRAFLLRLEGAKIGKNUV RPSVKLTVHIO LAIWARAFLAEVFVFDESSAPE EPPMAFDVG LAIWARTAIALIYKKYFLAAL SOVSKTFTO LAIWARTAIALIYKKYFLAAL SOVSTFTO ALLEAYRLGHVJMNIKKYFLAAL SOVSVTVFO ALLOXYRTHYJMNONKSIALATP ONOVSVTVFO ALLOXYRTHYJMNONKSIALATP ONOVSVTVFO ALLOXYRTHYJMNONKSIALATP ONOVSVTSO ALLOXYRTHYJMNOVKKYLAAL SOURFTTO ALLOXYRTHAMLMWKGRALATP ONOVSVSS ALLOXYRTHYJMNU SORLAVID SALLSTV ALLOXYRTHAMLONGVSVSTVELARL SOURFTTO ALLOXYRTHAMLONGVSVSVSS	UNPUALLER FILLER
VELGQGVSVGPFCHVQ AQLHDSVTVGPYSIIE ASIGANAHJGPFCIVG AVIGENVFJGPFCIIE AVIGANVFJGPFCVIG AVIGANVFJGPFCVIG AVIGANAHJGPFCIVG AVIDASTAKJGNVSIG AVIDASTAKJGNVSIG AVIFDSVLJGENVSIG	AUTDATIGRAVI1G- AUTDATIGRAV11G- AUTSPOATIGRAVS1G- NKS-PNLVVGYSYSG- VKN-PNLKVGYSYSG- VKN-PNLKVGYSYSG- VKN-PNLKVGRYSYSG- VKN-PNLKVGRYSYSG-	CEVYTTDI.ALAYQEGR- GEYYTTDI.ALAYQEGR- GEYYLTDL.ALAYQEGR- GEYYLTDL.ALANADDI- GEYYLPDVIEILANBGE- KRLRGKTLMYERNHSHP EQQEALVLI.EKFNQTSV- VFYRNCRJ.TRFPCYIRN- I.RKETKI.KHEYNTLSP- DR.RARQL.TRCYUNETP- ELLLTGAM.KRYNDTLG-	DLLAAGAWLKRYNSTLG- NAHFCSNKRKKNVLNN- NAIKVQLWMAVQATIF- ARCNAKKLHRRPNDTMGD DPARESYFEULLYLKG- DPANELWSTPLLYLKG- DPAVELWSTPLLYLKG- DPAVELWSTPLLYLKG- DPAVELWSTPLLYLKG- DPANELWSTPLLYLKG- DPANELWSTPLLYLKG- DPANELWSTPLLYLKG-	DPAARCHLTULTULTULTULTULTUL DPAARTTFEULTTYPG DPAARTTFEULTTYPG MBSPTTVPUANBEP- MBSPTTVPUANBEP-
LPXA_BRUAB LPXA_CHFU1 LPXA_ECOLI LPXA_BCOLI LPXA_BRONI LPXA_RICH LPXA_RICH LPXA_SALTY LPXA_SALTY LPXA_VEREN LPXD_HAEIN	LPXD_RICKI LPXD_SALTY LPXD_SALTY LPXD_SALTY CAT4_AGRTU CAT4_ECOLI CAT4_EOLI CAT4_ENTAE CAT4_ENTAE SATA_ENTFC VATA STAAU	WATA_BACSH MATA_BACSH GLMU_ECOLI GLMU_ECOLI GLMU_BECOLI THGA_ECOLI THGA_ECOLI VJV8_YEAST WBBJ_ECOLI YYAI BACSU YYAI BACSU NODL_RHILV	NODL_RHIME WCAF_BCOLI WCAF_BCOLI YA39_SCHPO CYSE_BCCUL CYSE_BCCUL CYSE_BCCUL CYSE_BCUL CYSE_BCUL CYSE_SALTY CYSE_SALTY CYSE_SALTY CYSE_SALTY	STRPH_SYNP7 STRPH_SYNP7 NIFP_AZOCH CCMM_STVP7 FBP_PSAE FBP_PSEAE TRDA_ECOLI 7304_METUA 7304_METUA 7304_METUA 7304_METUA DAPD_ATPL DAPD_HAEIN DAPD_HAEIN DAPD_HAEIN DAPD_HAEIN DAPD_HAEIN DAPD_HAEIN CAFG_STAAU

FIG. 3. Multiple alignment of the 58 sequences recovered from the similarity searches. This structure-based sequence alignment was obtained using the structural alignment of Fig. 2 as a seed (see text). Conserved positions are shaded using GENDOC with 90, 75, and 50% conserved, using PAM250 as scoring table. The Swiss-Protein ID is used to identify each sequence. See Table 1 for further reference.

LOVA DOUND .	-HYVTFSNNVMIGG	umol	Section 20		luopilpi	Buuk Place Salucia	TOYONATOU
LPAA_BRUAD :	-DHNVFANTATLAG	uiir	DUCE	GGGMA	HOPCR	CUTTO DOLLAR VS	DDVM ADDU
LPXA_CHRVI :	-NRCILANNAT AG	HVE	DIRA	SGQVA	HOPCK	AGVIG PU	PPIMLAD
LPXA_ECOLI :	- NCI ANNAT AG	HVS	DFA	GALL CMON	HOPUT	CARVINGGCSGVAQD	PPIVIAQUN
LPXA_HAEIN :	- DRCI ANNAT AG	HUE	DFV	GGMSA	HQFWI	GAHVMLOGGSMVSQD	PPIVMAQUN
LPAA_PROMI :	- RCI ANNGI GG	u c	IV	GGMSA	HOVAD	CEVEN OCCOURAGE	T DECLUCCK
LPXA_RICRI :	- NVVFANYVS AG - NRCI ANNAT AG	H G	DIDA	GGLSA	HQYAR	GEYSMIGGLSPVGAD	TPFGLVSSK
LPXA_SALTY :	-DRCILANNATLGG	HVS	DFA	GGMTA	HOPCI	GAHVMVGGCSGVAQD	PPIVIAQUN
LPXA_YEREN :	-DDTIIGNGVIIDN	OCOT NUMBER	DIFA	GGMLA	HOFCV	GAHVMVGGCSGVAQD	PPFVIAQUN
LPXD_ECOLI :	-DATIEDNVIIDN	QCQIAHNVV	MTA	AGGVI	AGSLK	GRICH GGASVINGH	EICDKVTVT
LPXD_HAEIN :	-ODTI KDLCRIDN	LCQIAHNUH	TGTA	AGGVI	AGST	GRICLIGGASVINGH	BICDRVTIT
LPAD_RICRI :	-DDTVIGNGVIIDN	LVQIGHGVK	KGS	VAQIG	AGSST	GRICA GGQVGIAGH	NIGDGAQVA
LPXD_SALTY :		QCQIANNVV	GUNTA	AGGVI	AGSIN	GRICHIGGASVINGH	EICDKVTVT
LPXD_TEREN :	- NTI GNGVI DN	QCQIAHNVV	NTA	AGGV1	AGSLK	GRYCHLOGASVINGH	REICORVIII
CAT4_AGRTU :	-SDVWIGSEANIMP	G W	GLGAV	GTRAL	TKD	EPYALVOGNPAKTIR	KRFDDDSIAL
CAT4_ECOLI :	-NDVWIGSEAM/MP	GIK	GIGAV	GSRAL	AKD	EPTITUGGNPAKSIR	KRESEEEISM
CAT4_ENTAE :	-NDVWIGSEAMVMP	GTK	GAV	GSRSL	TKD	EPIALVOGNPAKKIK	KRFTDEEISL
CAT4_MORMO :	-NDVWIGSEAMIMP	GIK	GLGAV	GSKEL	TKD	VPYALLOGSPAKQIK	KRFSDEEISL
SATA_ENTEC :	-NDVWIGKDVVIMP	GVK	GDGAL	AANSV	VKD	APYM A GNPANEIK	QRFDQDTINQ
VATA_STAAU :	-NDVWIGRDVTIMP	GVK	GEGAL	AAEAV	TKN	APYSINGGNPLKFIR	KRFSDGVIEE
MATA_BACSH :	-NDVWIGQNVTIMP	GW1	GDGAL	AANST	VKS	EPYSYSGNPAKFIK	KRFSDEKIEF
GLMU_ECOLI :	HRVKIGTGC	VIKNSV	G. DCE	SPYTVVEDAN	AAACT	GPFARERPGAELLEG	AHVGNFVEMK
GLMU_HAEIN :	DRVKIGTGC	VLKNWV	GNDVE	KPYSVLEDSI	GEKAA	GPFSRERPGAELAAE	THVGNFVEIK
GLMU_NEIGO :	DNVEIGANC	VIKNAK	CANSK	APFSHLEGCE	GENNR	GPYARLRPQARLADD	VHVGNFVEIK
GCAD_BACSU :	EDTIIGPHT	EIMNSA	GSRTM	K-QSVVNHSK	GND∨N	GPFAHIRPDSVIGNE	KIGNFVEIK
THGA_ECOLI :	RKNGEMYSF	PIT	GNVW	GSHVV	NPGVT	GDNSVIGAGSIVTKD	IPPNVVAAGV
THGA_LACLA :	RKRGAQYNK	K₩Y	E NVW	GAGVI	LPGVR	GKNSVIGAGSLVTKD	IPDNVVAFGT
WBBJ_ECOLI :	MRTLESS	AVV	GORVWI	GENVT	LPGTI	GNGVVVGANSVV-RG	FRKYCHCGV
YJV8_YEAST :	RLQGLEYAM	PVT	GENVW	GGGVS	IPGVN	GKNSVIAAGSVVIRD	I PENVVAAGN
MAA_ECOLI :	RNSGAELGK	PWT	GENVW	GGRAV	NPGVT	GDNVVVASGAVVTKD	PDNVVVG
YYAI_BACSU :	RKSGKEFGK	PWT	GDQVW	GGRAV	NPGVT	GDNAVIASGSVVTKD	PANTVVGGN
NODL_RHILV :	RQAGLQLGR	PVS	GRHAW	GGGAI	LPGVT	GDHAVIGAGSVVTRD	PAGSTAMON
NODL_RHIME :	RQAGLQL GR	PVR	GKHVW	GGGAI	LPGVT	GDHAVVGAGSVVTRD	PPGAKVMGS
WCAB_ECOLI :	ACPH		GNGVE	GANVI	LGDIT	GNNVTVGAGSVVLDS	VPDNALVVGE
WCAF_ECOLI :	ASQHFTIN	ATPIV	GEKCWI	ATDVF	APGVT	GDGTVVGARSSVFKS	PANVVCRGN
YA39_SCHPO :	LVQLLKHLTTNKQN	FIK	Q NVW	GMGVI	LPGVT	GEGSVIGAGAVVTKD	IPPNTLAVGS
CYSE BACSU :	GTGKEKG	KRHPT	KDAL	ATGAK	LGSIT	GEGSKIGAGSVULHD	VPDFSTVVGT
CYSE_BUCAP :	GTGSNTGK	NRHPI	RKNVT	GAGAK	LGNE	GQGVKVGAGSIVLKN	IPPFVTVVGV
CYSE ECOLI :	GTGKSGG	DRHPK	R GV	GAGAK	LGN	GRGAKIGAGSVVLOP	PPHTTAAGV
CYSE_HAEIN :	GTGKESG	DRHPK	R GV	GAGAK	LGNE	GKYAKIGANSVVLNP	PEYATAAGV
CYSE_HELPY :	GTGKFKG	KRHPT	GNRVVV	GAGAK	LGAIC	GDDVKIGANAVVLSD	PTGSTAVGS
CYSE_SALTY :	GTGKTSG	DRHPK	R GV	GAGAK	LGNE	GRGAKIGAGSVVLOP	VPPHTTAAGV
CYSE STAXY .	GTGKERG	KRHPD	CENUL	$\Delta = = = = = \Delta G = K$	LONT	NANUNTCANSUULNS	UDSVSTUUET
CYSE SYNP7 :	GTGKOSG	KRHPT	ANNVVV	GAGAK	LGN	GENVRIGAGSVVLRD	VPSDCTVVGI
CYSE_SYNY3 :	GTGKESG	KRHPT	GENVVV	GAGAK	LGNA	GDNVR IGAGSVVLRD	PADFTVVGV
SRPH_SYNP7 :	AKSFPRDETGAL	IKGOARHPV	EDVVI	YAGAT	LGRIT	GRGETIGGNVWLTRS	PAGSFISOA
NIFP AZOCH :	GTTGAKG	KRHPT	COVIN	GAGAK	LGPT	GAN RUGANSVVVOD	VPEGCTVVEI
CAH METTE :	-DDTFIGMQAFV-FK	SK	GNNCVI	EPRS-A	AIGUT	PDGRY PAGMVVTSO	AEADKLPEVT
CCMM SYNP7 :	-DDCFVGFRSTV-FN	AR	AGSVI	MMHA-	VODVE	PPGRY PSGAIITTO	OOADRLPEVR
FBP PSEAE :	-DYSLVGINAVILN	GAK	GKYCT	GANAL	PEGKE	PDGSLAMGSPGKVVR	ELSEPOKKMI.
YRDA ECOLT .	-NRVLUGMGSTLLD	GAT	FDVM	GAGSL	PONKR	ESCY VLGSP KOTR	PLSDEEKART.
CAIE ECOLI :	-RDALVGMNSVIMD	GAV	EST	AAMCE	KAGERO	EKROLMGTPAPAVRI	NVSDDELHWK
Y304 MET.IA .	-DNVLVGMNATILN	CAK	GENCT	GANAT	TONKE	PPNSIAL CVPCPU-V	RELTEETIKS
TARE PSEST .	-SGAV GMGVYL -GASTPI	DRASGEVRECE	PANA		1 QUALE	NPADDVI NPADDVI	DOUGI ACAUT
DAPD ACTPL .	- DOCUT SMOVET - COSTRIN	DRETGETHYCR	PACSIN	V		SCOT PSVC	CKYCI VCAUT
DAPD FCOLT	- CSV SMCVV -COSTRI	DRETGETHYCE	PACS			SCNI PEVIN	CEVELVENUT
DAPD HAFTN .	- GSV SMGVY - GQSTRIY - GCV SMGVF - GQSTKIY	DRETGETHYOP	PACSIN	V		SCOL POVO	CEVEL VCAUT
DAPD MYCBO .	- GSV SMGVY -GQSTRIN	DRETGETHYCE	PAGSIN	V		SON PSKC	COVGLUCAVI
CAPC STANU .	-NNVV/GAGSV/TK	DIGLIGEINIGR	DINUT	CNDAR	VTVCT	AVENUT END DYNU	NUNEVVIV
SHIG_SIAND :	THE ARONODA MALL		r Bran Th	GIVPAP	un nom	ALD K DEMOUTING	TATABULT I D

FIG. 3—Continued

present paper, we briefly describe each branch in the following paragraphs.

Branch A. This is composed of the enzymes LPXA and LPXD. Both are acyltransferases involved in the biosynthesis of lipid A. LPXA is a UDP-N-acetylglucosamine acyltransferase that catalyzes the first step. LPXD, UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase, catalyzes the third step (Vuorio *et al.*, 1991, 1994). LPXAs are grouped together into subbranch A1 and LPXDs into subbranch A2.

Branch B. This is composed of hexapeptide xenobiotic acetyltransferases that use acetyl-CoA to acylate a variety of compounds (Beaman *et al.*, 1998a). This branch is further divided into two clear subbranches: B1 is composed of chloramphenicol acetyltransferases, and B2 is composed of enzymes involved in the acetylation of other antibiotics. This branching pattern follows the functional classification of hexapeptide xenobiotic

acetyltransferases in two different functional groups (Beaman *et al.*, 1998a; Murray and Shaw, 1997).

Branch C. This is composed of UDP-N-acetylglucosamine pyrophosphorylases. This activity seems to be unrelated to the activities of the rest of the proteins studied. However, it should be noted that GLMU_ ECOLI, GLMU_HAEIN, and GLMU_NEIGO are bifunctional, additionally showing glucosamine-1-phosphate acetyltransferase activity (see Table 1 and Blattner *et al.*, 1997). Note that the region considered in this study, the L β H motif, is only part (about 170 amino acids in the carboxy-terminal region) of these rather long sequences (more than 400 amino acids).

Branch D. This includes a diverse group of enzymes that acetylate different types of sugars. Thus, we can find galactoside *O*-acetyltransferases participating in lactose biosynthesis (Hediger *et al.*, 1985), the maltose *O*-acetyltransferase (Blattner *et al.*, 1997), the nodula-

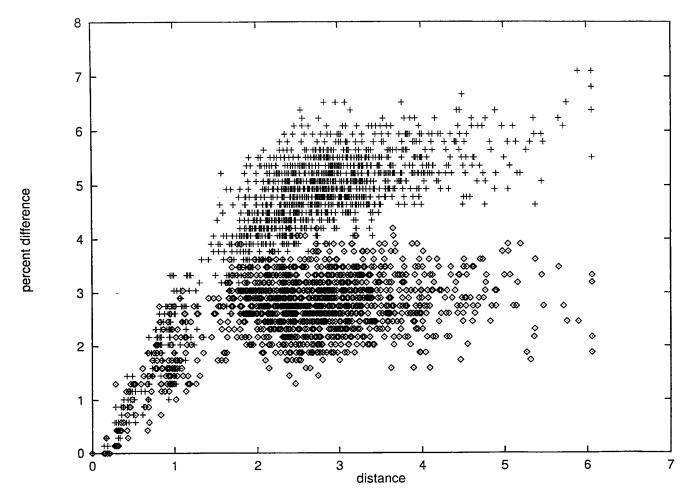


FIG. 4. For each pair of DNA sequences of Table 1, we plot the percentage of transitions (\diamond) and transversions (+) at second codon positions as a function of the corresponding distance. Note that for most sequence pairs transitions and transversions are saturated.

tion protein L involved in the O-acetylation of Nod factors (Baev and Kondorosi, 1992), colanic acid biosynthesis acetyltransferases (Stevenson *et al.*, 1996), and the putative galactoside *O*-acetyl transferase WBBJ_ECOLI, probably involved in lipopolysaccharide biosynthesis (Yao and Valvano, 1994).

Branch E. This is composed of serine acetyltransferases involved in cysteine biosynthesis. This is the only branch that differs between the two most-parsimonious trees. The two proteins that differ are NIFP_AZOCH and SRPH_SYNP7.

Branch F. This is the branch of main interest, containing γ -CA(CAH_METTE). According to the maximum-parsimony tree (Fig. 5, branch F), γ -CA and its closest relative, CCMM_SYNP7, share a common ancestor with the tetrahydrodipicolinate *N*-succinyltransferases, in disagreement with results of previous studies (Hewett-Emmett and Tashian, 1996). To obtain a more reliable topology, we performed a maximum-likelihood exhaustive search of branch F. In the tree obtained, shown in Fig. 6, CAH_METTE, CCMM_SYNP7, FBP_PSEAE, CAEI_ECOLI, YRDA_ECOLI, and Y304_METJA are clustered into subbranch F1, in agreement with Hewett-Emmett and Tashian (1996). This branch (F1) is clearly separated from the branch of tetrahydrodipicolinate N-succinyltransferases (F2).

DISCUSSION

We studied the evolutionary relationships between γ -CA and a very diverse group of proteins sharing the L β H sequence motif. We found that sequence alignments based on sequence alone are unreliable. The reason is, clearly, the nature of the sequences studied, which consist of the repetition of a short hexapeptide unit. To solve this problem we used a structural alignment of three members of the group with known crystallographic structures as a seed to obtain a reliable alignment.

The protein group studied is composed of very divergent sequences, which results in sequence similarities as low as 21%. However, the extremely low probability of the origin of hexapeptide repeated units by convergent evolution (Doolittle, 1994) strongly suggests that

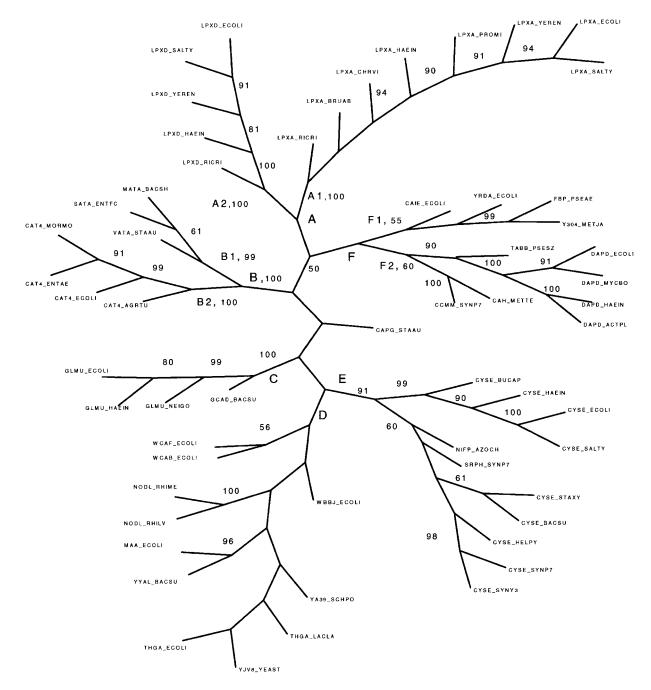


FIG. 5. Maximum-parsimony consensus tree. Topology obtained with the 50%-consensus rule for the two equally parsimonious trees of the sequences of Table 1. As a measure of node support, we indicate those bootstrap probabilities that are larger than 50%. Main branches and subbranches are identified by capital letters near nodes, for easier reference. The tree was displayed using TREEVIEW (Page, 1996).

we are dealing with a group of homologous proteins. Further support is obtained from the high structural similarity. We found a Sc of 5.4, which almost always suggests a functional and/or evolutionary relationship (Russell and Barton, 1992).

The most remarkable feature of the structural alignment (Figs. 1d and 2) and the structure-based alignment (Fig. 3) is that despite the high degree of divergence, key residues of the L β H sequence pattern are

highly conserved. From Fig. 3 we see that the aliphatic positions i are the most conserved feature of the motif. The glycines at positions i + 1 and the hydrophobic residues at positions i + 4 are also well conserved. The low degree of variation at these positions result from structural constraints, in agreement with the observation that this pattern is the sequential determinant of the L β H fold (Raetz, 1995; Kisker, 1994). On the other hand, loops disobey the repeated hexapeptide rule and

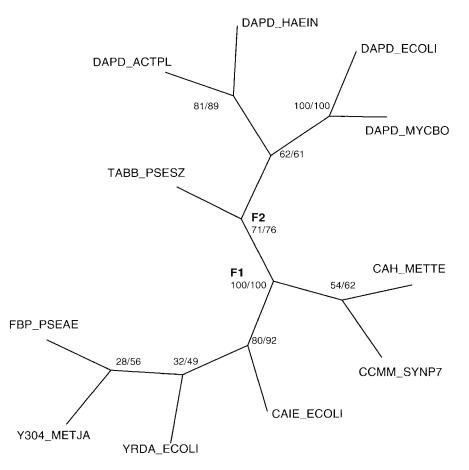


FIG. 6. Topology obtained from a maximum-likelihood exhaustive search of the sequences of the γ-CA main branch F of Fig. 5. Node support is indicated as RELL/RLS (see text). The tree was displayed using TREEVIEW (Page, 1996).

do not fold into the regular prism form. Loop and turn residues are less restricted to variation, so that the main differences between sequences are located in these segments. These structurally unconstrained regions hold most of the variation that led to the rather high functional diversity of this protein superfamily. Note, for example, that the histidines that participate in the active site of γ -CA (His 81, His 117, and His 122) and the residues of tetrahydrodipicolinate *N*-succinyltransferases that bind their substrates, such as Glu 169 among others (Beaman *et al.*, 1998b), are turn or loop residues.

We found that the high degree of divergence of the sequences studied results in the saturation of substitutions at the DNA level, so that the DNA does not contain enough phylogenetic signal to perform a meaningful phylogenetic analysis. Therefore, we based our evolutionary study on the protein alignment, since in such cases the use of protein methods improves phylogenetic inference (Kocher *et al.*, 1989).

The most remarkable feature of the phylogenetic tree obtained is that its main branches group together functionally related proteins (see Fig. 5 and Table 1), suggesting that this coarse topology resulted from functional diversification. Some of the main branches could be quite easily explained in terms of gene duplication events followed by functional specialization. The clearest example is branch A, in which enzymes are clustered by function rather than by species (see Fig. 5, branches A1 and A2). This pattern of duplication and/or speciation events, followed by functional specialization, can also be used to understand other main branches.

The branch containing γ -CA is more complex. In the first place, as we described in the previous section, maximum parsimony gives a wrong topology (Fig. 5, branch F), as judged by comparison with the more reliable topology obtained by a maximum-likelihood exhaustive search (Fig. 6) and with previous results (Hewett-Emmett and Tashian, 1996). The subbranch F1, composed of γ -CA's closest relatives (Fig. 6), has been studied with some detail before (Hewett-Emmett and Tashian, 1996). So far, very little is known about the function of the other members of this subbranch, as can be seen in Table 1. However, it is meaningful to note that the histidines of the active site of γ -CA, involved in the binding of Zn²⁺, are conserved in all the members of this subbranch. Therefore, the capacity to bind Zn^{2+} would have evolved before the divergence between γ -CA and its closest relatives. The other subbranch of the maximum-likelihood tree (Fig. 6, subbranch F2) is composed mainly of tetrahydrodipicolinate *N*-succinyltransferases. The active sites of these enzymes (Beaman *et al.*, 1998b) are clearly unrelated to those of subbranch F1. Although the γ -CA group (subbranch F1) and the tetrahydrodipicolinate *N*-succinyltransferases (subbranch F2) share a distant common ancestor, divergence was high enough to produce these two functionally unrelated groups.

A note of caution is in order before concluding. Even though we think that one can be quite confident in the coarse features of the phylogenetic tree reported, one should be much more careful when considering the details of each branch. In particular, we should note some troublesome aspects in the study of γ -CA evolution, such as the fact that CCMM_SYNP7, its closest relative, is a chimeric protein and the scarcity of information about the biological activity of other relatives. Further work is required to improve the evolutionary understanding of the relationship between γ -CA and its closest relatives.

In conclusion, we found that γ -CA belongs to a diverse superfamily of proteins that share the L β H domain and that are composed mainly of acyltransferases. The use of a structure-based protein alignment allowed us to perform a rather detailed evolutionary study. The coarse topology of the phylogenetic tree obtained can be rather easily understood in terms of functional diversification. Regarding the main branch containing γ -CA, we found that, in addition to the group of closest relatives, γ -CA is closely related to the tetrahydrodipicolinate N-succinyltransferases.

ACKNOWLEDGMENTS

We thank Arlin Stoltzfus, who kindly provided us with mol2con. We acknowledge the invaluable suggestions of the anonymous referees. This work was supported by the Universidad Nacional de Quilmes.

REFERENCES

- Adachi, J., and Hasegawa, M. (1994). MOLPHY: Programs for molecular phylogenetics, version 2.2. Distributed by the Institute of Statistical Mathematics, Tokyo.
- Alber, B., and Ferry, J. (1994). A carbonic anhydrase from the archaeon Methanosarcina thermophila. Proc. Natl. Acad. Sci. USA 91: 6909–6913.
- Altschul, S., Gish, W., Miller, E., Myers, W., and Lipman, D. (1990). Basic local alignment search tool. J. Mol. Biol. **215**: 403–410.
- Anderson, M., and Raetz, C. (1987). Biosynthesis of lipid A precursors in *Escherichia coli*: A cytoplasmic acyltransferase that converts UDP-*N*-acetylglucosamine to UDP-3-O-(*R*-3-hydroxymyristoyl)-*N*acetylglucosamine). *J. Mol. Biochem.* **262**: 5159–5169.
- Baev, N., and Kondorosi, A. (1992). Nucleotide sequence of the *Rhizobium meliloti* nodL gene located in locus n5 of the nod regulon. *Plant Mol. Biol.* 18: 843–846.
- Bairoch, A., and Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **20**(Suppl.): 2019–2022.

- Bairoch, A. (1993). The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucleic Acids Res.* 21: 3097–3103.
- Barton, G. (1990). Protein multiple sequence alignment and flexible pattern matching. *Methods Enzymol.* **183**: 403–428.
- Beaman, T., Binder, D., Blanchard, J., and Roderick, S. (1997). Three-dimensional structure of tetrahydrodipicolinate N-succinyltransferase. *Biochemistry* 36: 489–494.
- Beaman, T., Blanchard, J., and Roderick, S. (1998b). The conformational change and active site structure of tetrahydrodipicolinate *N*-succinyltransferase. *Biochemistry* 17: 10363–10369.
- Beaman, T., Sugantino, M., and Roderick, S. (1998a). Structure of the hexapeptide xenobiotic acetyltransferase from *Pseudomonas aeru*ginosa. Biochemistry **37**: 6689–6696.
- Benson, D., Boguski, M., Lipman, D., Ostell, J., and Ouellette, B. (1998). GenBank. Nucleic Acids Res. 26: 1–7.
- Blattner, F., Plunkett, G., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B., and Shao, Y. (1997). The complete genome sequence of *Escherichia coli*. *Science* 277: 1453–1474.
- Dicker, I., and Seetharam, S. (1992). What is known about the structure and function of the *Escherichia coli* protein FirA? *Mol. Microbiol.* 6: 817–823.
- Doolittle, R. (1994). Convergent evolution: The need to be explicit. *Trends Biochem. Sci.* **19**: 15–18.
- Downie, J. (1989). The nodL gene from *Rhizobium leguminosarum* is homologous to the acetyl transferases encoded by LacA and CysE. *Mol. Microbiol.* **3:** 1649–1651.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17: 368–376.
- Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, Univ. of Washington, Seattle.
- Gribskov, M., McLachlan, A., and Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* 84: 4355–4358.
- Hasegawa, M., and Kishino, H. (1994). Accuracies of the simple methods for estimating the bootstrap probability of a maximum likelihood tree. *Mol. Biol. Evol.* **11:** 142–145.
- Hediger, M., Johnson, D., Nierlich, D., and Zabin, I. (1985). DNA sequence of the lactose operon: The lacA gene and the transcriptional termination region. *Proc. Natl. Acad. Sci. USA* 82: 6414– 6418.
- Hewett-Emmett, D., and Tashian, R. (1996). Functional diversity, conservation and convergence in the evolution of the α -, β and γ -carbonic anhydrase gene families. *Mol. Phylogenet. Evol.* **5**: 50–77.
- Jermiin, L., Olsen, G., Mengersen, K., and Easteal, S. (1997). Majority-rule consensus of phylogenetic trees obtained by maximum likelihood analysis. *Mol. Biol. Evol.* 14: 1296–1302.
- Jones, D., Taylor, W., and Thornton, J. (1992). The rapid generation of mutation data matrices from protein sequences. CABIOS 8: 275– 282.
- Kanehisa, M. (1996). Toward pathway engineering: A new database of molecular pathways. Sci. Technol. Japan 59: 34–38.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 11–120.
- Kisker, C., Schindelin, H., Alber, B., Ferry, J., and Rees, D. (1996). A left-handed β-helix revealed by the crystal structure of a carbonic anhydrase from the archaeon *Methanosarcina thermophila*. *EMBO J.* **15**: 2323–2330.
- Kocher, T., Thomas, W., Meyer, A., Edwards, S., Pääbo, S., Villablanca, F., and Wilson, A. (1989). Dynamics of mitochondrial DNA

evolution in animals: Amplification and sequencing with conserved primers. *Proc. Natl. Acad. Sci. USA* 86: 6196–6200.

- Lüthy, R., Xenarios, I., and Bucher, P. (1994). Improving the sensitivity of the sequence profile method. *Protein Sci.* **3**: 139–146.
- McInerney, J. (1997). PUTGAPS, the Natural History Museum, London.
- Murray, I., and Shaw, W. (1997). O-Acetyltransferases for chloramphenicol and other natural products. Antimicrob. Agents Chemother. 41: 1–6.
- Neuwald, A. F., Liu, J., Lipman, D., and Lawrence, C. (1997). Extracting protein alignment models from the sequence database. *Nucleic Acids Res.* 25: 1665–1677.
- Nicholas, K., and Nicholas, H. (1997). Genedoc: A tool for editing and annotating multiple sequence alignments. Distributed by the authors.
- Page, R. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12: 357–358.
- Pearson, W., and Lipman, D. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85: 2444–2448.
- Raetz, C., and Roderick, S. (1995). A left-handed parallel β helix in the structure of UDP-N-acetylglucosamine acyltransferase. Science **270**: 997–1000.
- Roger, S., and Milner-White, J. (1995). RasMol: Biomolecular graphics for all. *Trends Biochem. Sci.* 20: 374.
- Russell, R., and Barton, G. (1992). Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and

residue confidence levels. Proteins: Struct. Funct. Genet. 14: 309–323.

- Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E., Jr., and Yunus, I. (1996). The metabolic pathway collection from EMP: The enzymes and metabolic pathways database. *Nucleic Acids Res.* 24: 26–28.
- Stevenson, G., Hobbs, M., Andrianopoulos, K., and Reeves, P. (1996). Organization of the *Escherichia coli* K-12 gene cluster responsible for production of the extracellular polysaccharide colanic acid. J. *Bacteriol.* 178: 4885–4893.
- Thompson, J., Higgins, D., and Gibson, T. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673–4680.
- Vuorio, R., Härkönen, T., Tolvanen, M., and Vaara, M. (1994). The novel hexapeptide motif found in the acyltransferases Lpxa and Lpxd of lipid A biosynthesis is conserved in various bacteria. *FEBS Lett.* 337: 289–292.
- Vuorio, R., Hirvas, L., and Vaara, M. (1991). The Ssc protein of enteric bacteria has significant homology to the acyltransferase Lpxa of lipid A biosynthesis, and to three acetyltransferases. *FEBS Lett.* 292: 90–94.
- Yao, Z., and Valvano, M. (1994). Genetic analysis of the O-specific lipopolysaccharide biosynthesis region (rfb) of *Escherichia coli* K-12 W3110: Identification of genes that confer group 6 specificity to *Shigella flexneri* serotypes Y and 4a. J. Bacteriol. **176**: 4133– 4143.