Phylogenetic Analysis of Carbamoylphosphate Synthetase Genes: Complex Evolutionary History Includes an Internal Duplication Within a Gene Which Can Root the Tree of Life

Fiona S. Lawson, Robert L. Charlebois,* and Jo-Anne R. Dillon*†*

*Department of Biology and †Department of Microbiology and Immunology, University of Ottawa, Ontario, Canada

Carbamoylphosphate synthetase (CPS) catalyzes the first committed step in pyrimidine biosynthesis, arginine biosynthesis, or the urea cycle. Organisms may contain either one generalized or two specific CPS enzymes, and these enzymes may be heterodimeric (encoded by linked or unlinked genes), monomeric, or part of a multifunctional protein. In order to help elucidate the evolution of CPS, we have performed a comprehensive phylogenetic analysis using the 21 available complete CPS sequences, including a sequence from *Sulfolobus solfataricus* P2 which we report in this paper. This is the first report of a complete CPS gene sequence from an archaeon, and sequence analysis suggests that it encodes an enzyme similar to heterodimeric CPSII. We confirm that internal similarity within the synthetase domain of CPS is the result of an ancient gene duplication that preceded the divergence of the Bacteria, Archaea, and Eukarya, and use this internal duplication in phylogenetic tree construction to root the tree of life. Our analysis indicates with high confidence that this archaeal sequence is more closely related to those of Eukarya than to those of Bacteria. In addition to this ancient duplication which created the synthetase domain, our phylogenetic analysis reveals a complex history of further gene duplications, fusions, and other events which have played an integral part in the evolution of CPS.

Introduction

Carbamoylphosphate synthetase (CPS) catalyzes the formation of carbamoylphosphate from CO_2, ATP, and ammonia or glutamine, for pyrimidine biosynthesis, arginine biosynthesis, or the urea cycle. CPSs are currently classified into three groups: CPSII (E.C. 6.3.5.5), a glutamine-dependent CPS, catalyzes the first committed step in pyrimidine biosynthesis and is also critical for arginine biosynthesis in bacteria and fungi. CPSI (E.C. 6.3.4.16) utilizes ammonia rather than glutamine and is activated by acetylglutamate (Marshall 1976; Anderson 1980). It is present in ureotelic vertebrates, where the arginine biosynthetic pathway functions as the urea cycle and the CPS is used to harvest ammonia. CPSIII, the most recently discovered CPS, is found in fish and in some invertebrates. CPSIII is acetylglutamate-activated and, although it is related to CPSI, it is glutamine-dependent (Hong et al. 1994).

All CPS enzymes comprise an amidotransferase domain and synthetase domain. The amidotransferase domain binds ammonia or glutamine, transferring the amide group to the synthetase domain which completes the reaction, including two phosphorylation steps. There is significant internal similarity within the synthetase domain of CPS from all organisms examined, the result of a proposed ancient duplication of a kinase gene (Nyunoya and Lusty 1983). All bacteria studied to date contain a heterodimeric form of the enzyme, which comprises a 40-kDa amidotransferase subunit and a 120-kDa synthetase subunit, while most eukaryotes examined contain a monomeric CPS which corresponds to a proposed fusion of these two domains (Nyunoya, Broglie, and Lusty 1985). The heterodimeric CPS can be encoded by genes which are co-transcribed, separately transcribed, or even on separate chromosomes (Werner, Feller, and Piérard 1985; Kwon et al. 1994; Lawson, Billowes, and Dillon 1995). Significant variation is seen within the intervening sequence between the CPS genes encoding the heterodimeric form of the enzyme, most notably between very closely related species of Proteobacteria, such as *Neisseria gonorrhoeae* and *Neisseria meningitidis*, though the genes themselves are highly conserved (Kwon et al. 1994; Lawson, Billowes, and Dillon 1995).

All Proteobacteria studied use one CPS enzyme for both arginine and pyrimidine biosynthesis while the Gram-positive bacteria examined have two CPS enzymes, which are separately regulated for arginine and pyrimidine biosynthesis (Paulus and Switzer 1979). Two CPS enzymes have also been identified in yeast and vertebrates, one involved in the pyrimidine biosynthetic pathway and one used either for arginine biosynthesis or in the urea cycle (Hong et al. 1994). However, recent reports show that two apicomplexan protozoans contain only one CPS enzyme (Flores, O'Sullivan, and Stewart 1994; Chansiri and Bagnara 1995).

To elucidate the complex evolution of CPS, we present a comprehensive phylogenetic analysis of all completely sequenced CPS genes, with a focus on recently obtained sequences including the first complete archaeal CPS genes, reported in this paper. The internal similarity present in the synthetase domain of CPS permits a rooting of the tree of life. Previously, the root has been deduced using elongation factors (Iwabe et al. 1989), ATPases (Gogarten et al. 1989), and aminoacyl-tRNA synthetase genes (Brown and Doolittle 1995); however, there is controversy regarding the use of ATPases and elongation factors, mostly due to concerns

Abbreviation: CPS, carbamoylphosphate synthetase.

Key words: carbamoylphosphate synthetase, arginine biosynthesis, pyrimidine biosynthesis, phylogenetic trees, tree of life, rooted trees, *Neisseria gonorrhoeae*, *Sulfolobus solfataricus* P2.

Address for correspondence and reprints: Dr. Jo-Anne R. Dillon, Professor and Chair, Department of Microbiology and Immunology, University of Ottawa, 45 1 Smyth Road, Ottawa, Ontario, Canada, K1H 8M5. E-mail: jdillon@labsun1.med.uottawa.ca.

Mol. *Biol. Evol.*13(7):970–977. 1996
 ISSN: 0737-4038

about paralogy versus orthology, lateral gene transfer, or a lack of statistical robustness (Forterre et al. 1993; Hilario and Gogarten 1993; Creti et al. 1994). Phylogenetic analysis using CPS genes avoids some of these problems and represents the first use of a metabolic gene to root the tree of life.

Materials and Methods

DNA Sequences

We previously cloned and sequenced the CPS genes from N. *gonorrhoeae* CH811(Picard and Dillon 1989; Lawson, Billowes, and Dillon 1995). *Pseudomonas stutzeri* and *Pseudomonas aeruginosa* synthetase sequences were kindly supplied by Drs. A. T. Abdelal and C.-D. Lu (Georgia University, Atlanta, Ga.) before publication or submission to GenBank. The sequence of the *Sulfolobus solfataricus P2* carbamoylphosphate synthetase genes, which we report in this paper, was obtained as part of a larger effort to sequence the entire genome of this crenarchaeote (Sensen et al. 1996). Random plasmid libraries of nebulized cosmid DNA in pUC 18 were sequenced in both orientations using an ABI 373A automated sequencer, to a coverage redundancy of about three. Oligonucleotide primers were then synthesized for directed sequence finishing at the cosmid DNA level to link contigs, to fill single-stranded gaps, and to resolve ambiguities. The entire sequence of the cosmid (sh02a07.08) containing *carAB* will be published elsewhere. The sequence reported in this paper, comprising only the CPS genes from S. *solfataricus P2,* has been submitted to GenBank, accession number U33768.

All other sequences were obtained by screening the GenBank and EMBL databases (as of August 1995) for all sequences reported to contain complete CPS genes. Sequences from the following organisms were used, with accession numbers in parentheses. Note that for organisms in which there are two CPS enzymes, the particular CPS gene sequence obtained is designated as either CPSI, CPSIII, Arg (CPSII-arginine-specific), or Pyr (CPSII-pyrimidine-specific). For organisms in which there is only one CPS enzyme, no such designation is given: *Escherichia coli* (J01597), *Salmonella typhimurium* (X13200), *P. stutzeri* (U04993), *P. aeruginosa* (U04992), *N. gonorrhoeae* (U11295), *Bacillus caldolyticus* (Pyr; X73308), *Bacillus subtilis* (Arg and Pyr; 226919, M59757), S. *solfataricus* (U33768), *Plasmodium falciparum* (L32150), *Babesia bovis* (U18792), *Saccharomyces cerevisiae* (Arg and Pyr; K02 132, K01178, M27 174), *Neurospora crassa* (Arg; JO55 12), *Trichosporin cutaneum* (Arg; L08965), *Dictyostelium discoideum* (Pyr; X14533, X55433), *Squalus acanthias* or spiny dogfish shark (CPSIII; L31362), *Rana catesbeiana* or bullfrog (CPSI; U05193), Syrian hamster (Pyr; J05503), rat (CPSI; M11710, M12318–28), and human (CPSI; D90282). Note that only the amidotransferase gene has been sequenced from N. *crassa* and S. *typhimurium* (Davis, Ristow, and Hanson 1980; Kilstrup et al. 1988) and only the synthetase gene has been sequenced from *T. cutaneum* (Reiser et al. 1994). For organisms with two CPS enzymes, S. *cerevisiae* and *B. subtilis are* the only organisms from which both CPS gene sequences have been obtained. The sequence from *Drosophila melanogaster* was not included because of possible mistakes noted by Simmer et al. (1990). No partial sequences were included in our study.

Alignment of Sequences

Deduced protein sequences from the DNA sequences obtained were aligned using the Clustal multiple-sequence alignment method of GDE version 2.2 (Steve Smith, unpublished), using the identity matrix or PAM 250 matrix with a fixed gap penalty of 40 and a floating gap penalty of 10. After initial alignment of the sequences, the central amidotransferase and synthetase domains were determined and used for subsequent phylogenetic analysis. These domains correspond to residues 38 to 367 of the deduced protein from the *E. coli carA* sequence (amidotransferase domain) and residues 11 to 9 16 of the deduced protein from *E. coli carB* (synthetase domain). The synthetase domain has internal similarity (Nyunoya and Lusty 1983), and so additional alignments were also performed using the two similar regions of this synthetase domain. These regions correspond to residues 11 to 333 in the N-terminus, and residues 563 to 875 in the C-terminus, of the deduced protein from *E. coli carB.* All protein alignments generated were used as a template to produce a corresponding alignment of the DNA sequences.

Phylogenetic Tree Construction

Phylogenetic trees were produced using PHYLIP (Phylogeny Inference Package) version 3.5 (Felsenstein 1993). Parsimonious trees were compiled using the dnapars and protpars methods of PHYLIP. Distance matrices were developed using dnadist with the Kimura two-parameter model (Kimura 1980), and protdist using the categories model of Hall (Felsenstein 1993). Trees were constructed from distance matrices using Fitch, Kitsch, and neighbor-joining methods. All bootstrap analysis was done using 100 multiple data sets with a jumbling factor of 1 (J = 1). A jumbling factor of 50 was used when no bootstrapping was performed. The maximum-likelihood method was also employed, using Molphy (maximum-likelihood inference of protein phylogeny) version 2.1.2 (protml version 1) with automatic default settings (Adachi and Hasegawa 1993).

Results and Discussion

Sulfolobus solfataricus CPS is Similar to Other Heterodimeric CPSII Enzymes

We report the first complete sequence of CPS genes from an archaeon, S. *solfataricus P2*. The DNA sequence analysis shows that this archaeal CPS is heterodimeric, encoded by genes which overlap by 4 bp and are present in the order *carA–carB*. Immediately upstream of *carA are* genes encoding the final two steps of arginine biosynthesis (*argG* and *argH*) in an arrangement suggesting co-transcription with *carAB* (unpublished data). Further analysis of nearby gene organization and transcriptional signals will be published elsewhere.

These *carA* and *carB* genes, which are of sizes 1,101 and 3,153 bp, respectively, are similar in size to those encoding other heterodimeric CPSs. An amidotransferase subunit of M_r 41,480 and a synthetase subunit of M_r 118,204 are predicted. This is notable since Legrain et al. (1995) recently characterized the CPS enzyme from the archaeon *Pyrococcus furiosus*, and found it to be atypical in size (M_r 70,000). The deduced amino acid sequence for the amidotransferase domain and the synthetase domain of the S. *solfataricus* CPS sequence is shown in figure 1. This sequence shows approximately 37% and 43% identity with *E. coli* and bullfrog (CPSI) sequences, respectively (for reference, *E. coli* and bullfrog CPS share 37% identity, and *E. coli* and N. *gonorrhoeae* CPS share 67% identity). Analysis of the protein sequence shows that it contains all the conserved domains found in other CPS enzymes. Notably, the S. *solfataricus* CPS enzyme contains internal similarity between the first and second thirds of the synthetase domain, as has been observed in all other organisms, including the archaeon *Methanosarcina barkeri* (Schofield 1993).

As expected, residues thought to be involved in ATP binding, and found in the synthetase domain of all CPSs (Post, Post, and Raushel 1990), are present in the S. *solfataricus* CPS (fig. 1). Within the amidotransferase domain, a cysteine residue plus other residues which have been implicated in glutamine binding for CPSII (Rubino, Nyunoya, and Lusty 1986; Miran, Chang, and Raushel 1991) are present, suggesting that this enzyme is most similar to other glutamine-dependent CPSII enzymes (fig. 1). Within the synthetase domain, residues implicated in acetylglutamate binding, which are found only in CPSI and CPSIII enzymes (Geschwill and Lumper 1989), are not present. However, no definite conclusion can be made at this point regarding the ability of the S. *solfataricus* CPS enzyme to bind glutamine, especially since recent findings (the CPSI of bullfrog contains the cysteine residue known to bind glutamine and yet is ammonia-dependent) have shown that other as yet unidentified residues must also play a part (Helbing and Atkinson 1994). The atypically sized archaeal CPS of *P. furiosus* uses ammonia and not glutamine as its nitrogen donor (LeGrain et al. 1995). However, the residues present in the S. *solfataricus* CPS are consistent with a glutamine-dependent CPSII-type enzyme.

Phylogenetic Analysis of CPS Genes

Forty phylogenetic trees were constructed using five different methods to analyze both protein and DNA sequences of four different alignments (complete CPS sequences, just the amidotransferase domain, just the synthetase domain, and the duplication within the synthetase domain). Trees constructed using complete CPS sequences (i.e., combined amidotransferase and synthetase domains) showed high confidence in branching order for almost all organisms, with 9 out of 14 nodes consistently having bootstrap values of 100 out of 100 replicates. An example of one tree, constructed using the Fitch distance matrix method with protein sequences, is shown in figure 2. In this tree, the sequences grouped into six clusters, comprising the Proteobacteria

A

MKLENKKGYLYLEDGTFIEGYSFGAKGIKVGEWFTTSMNG
YVESLTDPSYKGQILIITHPLVGNYGVPEKKYEQGILTNFE
SERIQVEGLIVAEHTYPSKWNSALTLDEWLKSENVPGVFDV
DTRMIVKKIRTYGTMMGIIASELEIDDPRKYLEKKYDEIDF
TQFTSPKSPIFHPNTGDMIWVDCGIKHGILYGLYKRGFSI
VRVPCSFSASKIIEYNPKGIVFSNGPGNPNLLENQIKTFSE
LVEYKIPILGI~LGHQIATLALGGKIKKMKFGHRAINKPVI
ESNSNKCYISTHNHGYGIISKNDIPPNTKIWFYNPDDYTIE
GLIHEKLPIITTQFHPEARPGPWDTTWVFDKFRTMVTGK

B

MRETPKKVLVIGSGPIKIAEAAEFDYSGSQALKALKEEGIE
TVLVNSNVATVQTSKKFADKLYMLPWWWAVEKVIEKERPD
GIMIGFGGQTALNVGVDLHKKGVLQKYNVKVLGTQIDGIEK
ALSREKFRETMIENNLPVPPSLSARSEEEAIKNAKIVGYPV
MVRVSFNLGGRGSMVAWTEEDLKKNIRRALSQSYIGEVLLE
KYLYHWIELEYEVMRDKKGNSAVIACIENLDPMGVHTGEST
WAPCQTLDNLEYQNMRTYTIEVARSINLIGECNVQFALNP
RGYEYYIIETNPRMSRSSALASKATGYPLAYVSAKLALGYE
LHEVINKVSGRTCACFEPSLDYIVTKIPRWDLSKFENVDQS
LATEMMSVGEVMSIGRSFEESLQKAIRMLDIGEPGWGGKV
YESNMSKEEALKYLKERRPYWFLYAAKAFKEGATINEVYEV
TGINEFFLNKIKGLVDFYETLRKLKEIDKETLKLAKKLGFS
DEQISKALNKSTEYVRKIRYETNTIPWKLIDTLAGEWPAV
TNYMYLTYNGTEDDIEFSQGNKLLIIGAGGFRIGVSVEFDW
SWSLMEAGSKYFDEVAVLNYNPETVSTDWDIARKLYFDEI
SVERVLDLIKKEKFRYVATFSGGQIGNSIAKELEENGVRLL
GTSGSSVDIAENREKFSKLLDKLGISQPDWISATSLGEIKK
FANE*VGFPVLVRPSYVLSGGSSMKI*AYSEEELYEYVRRATEI
SPKYPWISKYIENAIEAEIDGVSDGNKVLGITLEHIEKAG
VHSGDATMSIPFRKLSENNVNRMRENVLNIARELNIKGPFN
VQFWKENTPYIIELN **LRASRSMPFSSKAKGINLINESMKA**
IFDGLDFSEDYYEPPSKYWAVKSAQFSWSQLRGAYPFLGPE
MKSTGEAASFGVTFYDALLKSWLSSMPNRIPNKNGIALVYG
NKNLDYLKDTADNLTRFGLTVYSISELPLQDIETIDKMKAE
ELVRAKKVEIIVTDGYLKKFDYNIRRTAVDYNIPIILNGRL
GYEVSKAFLNYDSLTFFEISEYGGGI

FIG. I.-Deduced amino acid sequence of the CPS subunits encoded by *S. solfataricus* P2 *carA* **(A)** and *carB* **(B).** Residues implicated in glutamine binding and found in all glutamine-dependent CPS enzymes (CPSII) are marked in double-underlined, bold text. Regions proposed to contain the two ATP binding sites are shown in single-underlined, bold text, and regions with similarity to glycine-rich loops found important for catalysis in other ATP-binding proteins are highlighted in single-underlined, italic, bold text. The two cysteine residues conserved in acetylglutamate-dependent CPSs (CPSI, CPSIII) are not present in this archaeal CPS sequence: residues in their corresponding location are in bold text and boxed.

(P. stutzeri, P. aeruginosa, E. coli, N. gonorrhoeae), Gram-positive bacteria (B. *caldolyticus, B. subtilis),* Archaea (S. *solfaturicus),* apicomplexan protozoans *(P. falciparum, B. bovis),* eukaryotic CPSII arginine- and pyrimidine-specific enzymes (S. *cerevisiae* Arg, S. *cerevisiae* Pyr, *D. discoideum* Pyr), and eukaryotic CPSI and CPSIII enzymes (S. *acanthias* CPSIII, bullfrog CPSI, human CPSI). Similar branching order was noted in all trees constructed, with the following minor discrepancies: the *B. subtilis* arginine-specific (Arg) CPS some-

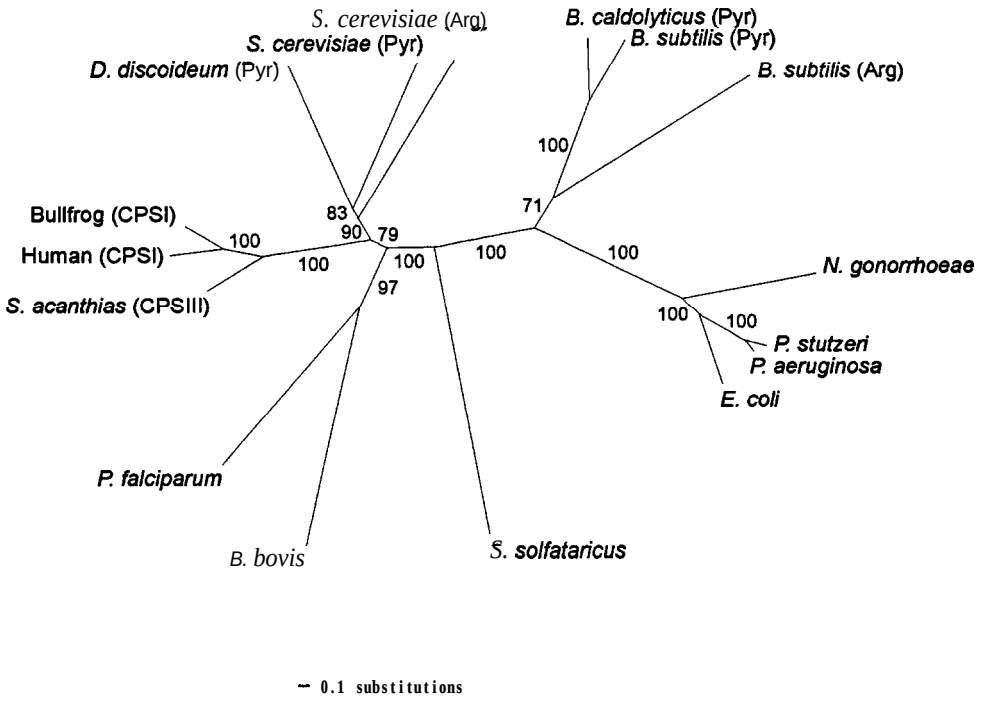

FIG. 2.-Phylogenetic tree constructed from alignments of deduced amino acid sequences from complete CPS genes, using the Fitch distance matrix method. Bootstrap values (percentage out of 100 replicates) are shown at each node with the scale for branch lengths shown below the figure. For organisms which contain two CPS enzymes, the sequence is identified by the letters "Arg" (for arginine-specific CPSII), "Pyr" (for pyrimidine-specific CPSII), "CPSI," and "CPSIII."

times formed its own lineage, the protozoan sequences would not cluster together when the Kitsch distance matrix method was used (a method which imposes equal divergence on all lineages), and some variability in branching order was observed within the metabolically specialized (i.e., arginine-, pyrimidine-, or urea-cycle-specific) eukaryotic sequences. When trees were constructed using just the amidotransferase domain of CPS, other sequences could be included, such as the N. *crassa* (Arg) sequence and *T. cutaneum* (Arg) sequence, which clustered with *S. cerevisiae* (Arg) as a monophyletic group. Similarly, S. *typhimurium* clustered with *E. coli*, and rat (CPSI) grouped with human (CPSI), as expected.

Phylogenetic Trees Constructed Using the Internal Duplication Within the Synthetase Domain of CPS: Rooting the Tree of Life Groups an Archaeon with Eukaryotes

The internal similarity observed within the first and second thirds of the synthetase domain of all CPS sequences was used to construct phylogenetic trees as shown in figure 3. Sequences corresponding to the first half of the duplication clustered together as one monophyletic group and the sequences corresponding to the second half of this region formed a similar cluster. This confirms that the duplication that formed the ancestral synthetase domain preceded divergence of the Bacteria, Eukarya, and Archaea (Schofield 1993), and allows this duplication to be used to root the tree of life.

Within all trees constructed using this duplication, the archaeal sequence grouped with high confidence with the eukaryotes, and always as the deepest branch in the clade. This phylogenetic analysis therefore shows that this archaeal sequence is more related to that of the Eukarya than to that of the Bacteria. The ability of the internal duplication to produce trees well resolved in the deepest branches argues that CPS may be useful in addressing the issue of archaeal monophyly (Lake et al. 1984) as well.

This is the first use of a gene involved in a metabolic pathway to root the tree of life, and so represents an important confirmation of these ancient phylogenetic relationships. Previously, the root had been deduced using elongation factors (Iwabe et al. 1989), ATPases (Gogarten et al. 1989), and aminoacyl-tRNA synthetase genes (Brown and Doolittle 1995); however, there has been concern that the full family of ATPase genes has not yet been clearly determined, and that horizontal gene transfer cannot be discounted (Forterre et al. 1993; Hilario and Gogarten 1993). Also, the statistical reliability of the trees produced using the elongation factors has been criticized (Creti et al. 1994). Our phylogenetic trees, however, are rooted by a duplication which occurred not to form separate genes, but to form one gene. This means that we can be fairly certain that we are comparing orthologous sequences. We found no evidence indicating horizontal transfer of CPS genes, branching order within our trees being consistent with those constructed using other gene sequences.

Gene Fusion Events in Eukaryotic Carbamoylphosphate Synthetase

The general organization of CPS genes, not including introns, is shown schematically in figure 4 alongside

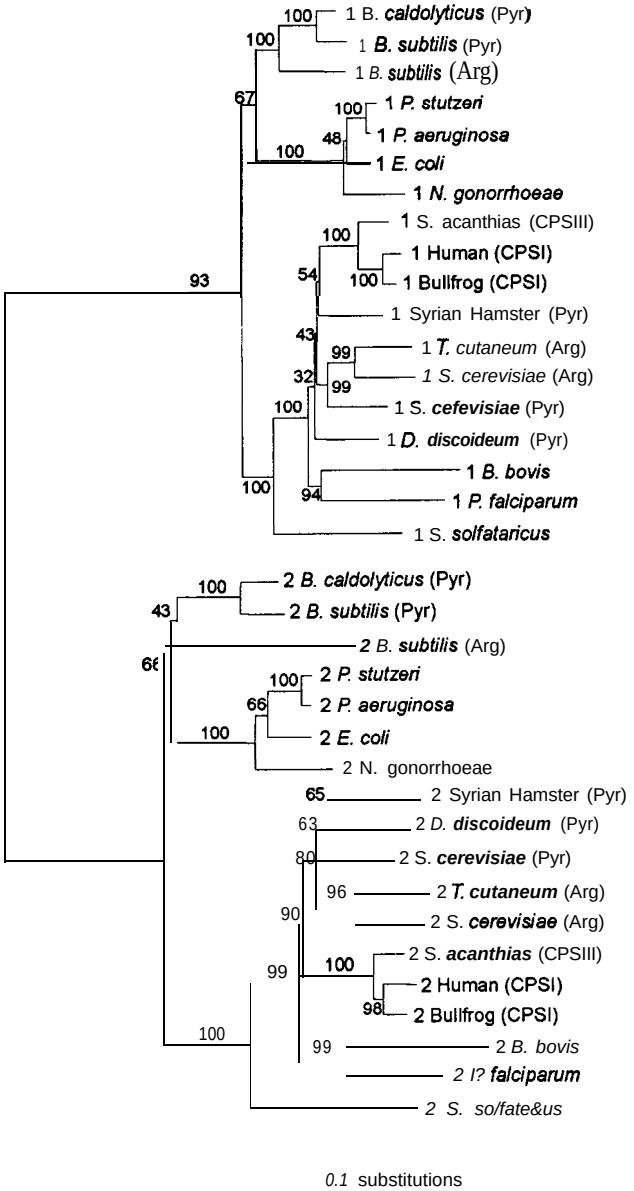

FIG. 3.—Phylogenetic tree constructed from alignments of the deduced protein sequence of the first and second thirds of the synthetase domain of CPSs, using the neighbor-joining distance matrix method. Bootstrap values (percentage out of 100 replicates) are shown at each node with the scale for branch lengths shown below the figure. The number preceding the name of each organism indicates whether that sequence was derived from (1) the first third of the synthetase domain or (2) the second third of the synthetase domain. For organisms which contain two CPS enzymes, the sequence is further identified by the letters "Arg" (for arginine-specific CPSII), "Pyr" (for pyrimidine-specific CPSII), "CPSI," and "CPSIII." Note that the branching orders determined from both thirds of the synthetase domain are similar, with the exception that the eukaryotic "Pyr" sequences cluster together in one case (second third of the synthetase domain) and in the other case (first third) they branch separately from one another. This branching order determined from the first third of the synthetase domain is not observed using any other method of tree construction, although the branching order for these eukaryotic "Pyr" sequences still varies.

one of the phylogenetic trees constructed. The phylogenetic tree suggests that the ancestral CPS enzymes were heterodimeric, since the deepest branches in the tree represent organisms with separate genes encoding the amidotransferase and synthetase domains. These genes are always found linked in the same order, suggesting they were closely linked in the progenote. They apparently fused early in the history of the eukaryotic lineage, prior to the origin of the apicomplexan protozoa, since the apicomplexan CPS is monomeric and branches earliest within the eukaryotes. Recently, van den Hoff et al. (1995) concluded that separate gene fusion events led to the monomeric eukaryotic enzymes CPSI and CPSII, since fungal CPSII (Arg) is heterodimeric (Lacroute et al. 1965; Davis, Ristow, and Hanson 1980). However, in light of these new findings regarding the apicomplexans (Flores, O'Sullivan, and Stewart 1994; Chansiri and Bagnara 1995), a more likely hypothesis is that a single gene fusion event occurred in eukaryotes, with the fungal heterodimeric CPSII (Arg) arising from a subsequent redivision of these genes.

The apicomplexan protozoan CPS has the unusual feature of containing large translated insertions between functional domains of the enzyme (Flores, O'Sullivan, and Stewart 1994; Chansiri and Bagnara 1995). This is likely a reflection of these particular organisms, which are noted for having large polypeptide insertions between functional domains of proteins (Flores, O'Sullivan, and Stewart 1994). The pyrimidine-specific CPS enzymes of eukaryotes have fused with other enzymes of the pyrimidine biosynthetic pathway (Davidson et al. 1993) and, based on the branching order observed in most trees constructed, this is likely due to a gene fusion event which occurred after the duplication of the CPS genes in eukaryotes.

Duplication and Divergence of CPS Genes in the Gram-Positive Bacteria and in the Eukaryotes

In all phylogenetic trees constructed, the two CPS enzymes present in the Gram-positive bacteria and in the Eukarya did not cluster together (figs. 2–4), confirming a recent report by van den Hoff et al. (1995) that separate gene duplication events led to the formation of two CPS enzymes in each of these lineages.

In the Gram-positive bacteria, the duplication leading to the formation of two CPS enzymes apparently occurred after the divergence of the Gram-positive bacteria from the Proteobacteria, since the Gram-positive enzymes are more related to each other than either is to the proteobacterial sequence. Only two Gram-positive bacteria have been examined so far, both of the genus *Bacillus,* and so it will be interesting to see if all Gram-positive bacteria have two enzymes, or if this feature is limited to certain genera and is thus a recent event. Preliminary findings based on studies of auxotrophic mutants indicate that *Lactobacillus* spp. have two CPS enzymes as well (Bringel 1994).

The single CPS present in apicomplexan protozoans, branching deepest within the eukaryotic tree, suggests that the CPS duplication within the eukaryotes occurred after their divergence. Van den Hoff et al. (1995)

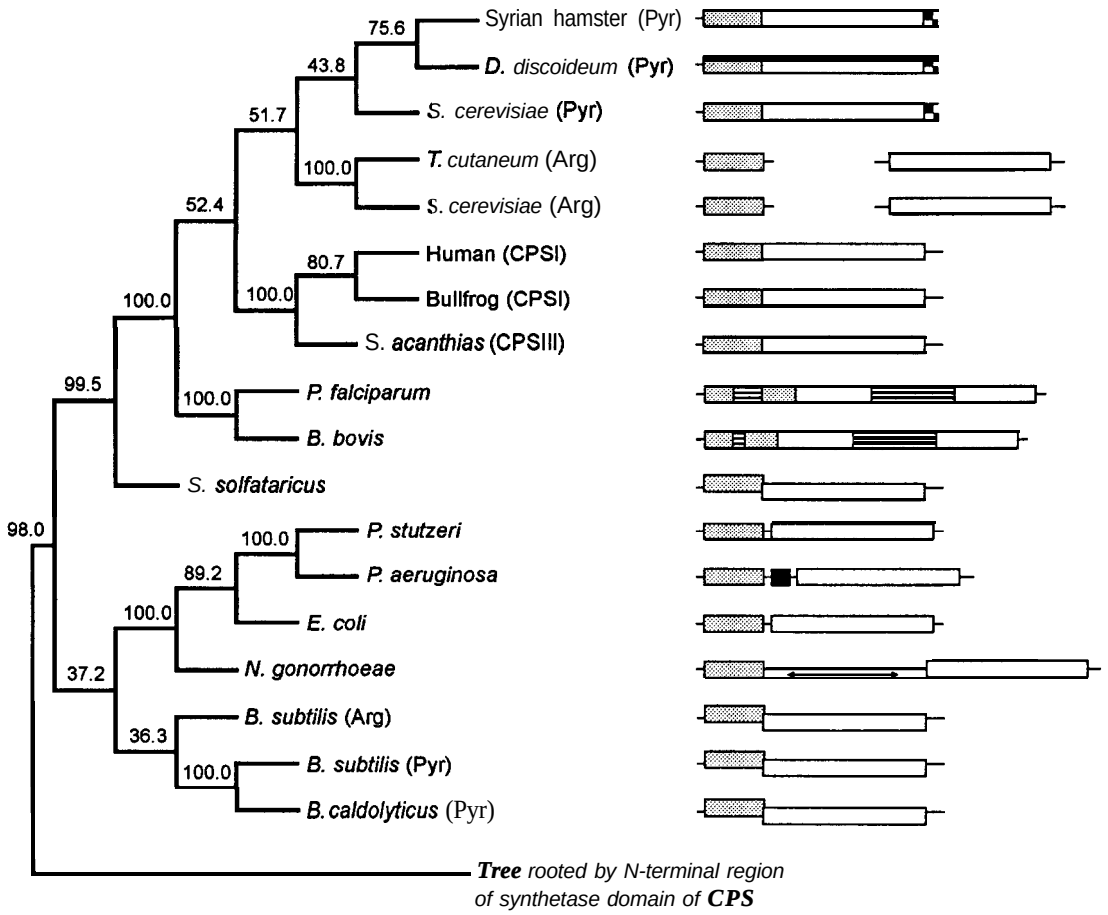

FIG. 4.-Phylogenetic tree constructed from alignments of the deduced protein sequence of the first and second thirds of the synthetase domain of CPSs, using the method of maximum parsimony. Only the branching order determined from the second third of the protein is shown, with the first third used as a root. Bootstrap values (out of 100 replicates) are shown at each node. For organisms which contain two CPS enzymes, sequences are further identified by the letters "Arg" (arginine-specific CPSII), "Pyr" (pyrimidine-specific CPSII), "CPSI," and "CPSIII." The organization of the CPS genes in each organism is also shown schematically, with regions encoding the common amidotransferase and synthetase domains of CPS shown as shaded and white boxes, respectively. Introns are not included. Note that the *S. cerevisiae* (Arg) and *T. cutaneum* (Arg) amidotransferase and synthetase domains are encoded by unlinked genes, and that the corresponding genes in *S. solfataricus* and *Bacillus* species overlap. Checkered boxes denote a portion of the aspartate transcarbamylase domain which is fused to the pyrimidine-specific CPS genes of eukaryotes. Striped boxes denote large translated sequences which are present between functional domains of the CPS of *P. falciparum* and *B. bovis*. The black box represents an unidentified open reading frame found between the *P. aeruginosa* CPS genes, The arrow denotes a variable intergenic sequence present in isolates of N. *gonorrhoeae*. Note that the N. *gonorrhoeae* CPS genes are separately transcribed, while all other bacterial CPS genes are part of operons.

proposed that this duplication of CPS genes occurred between the branching off of plants and fungi, since the fungi and animals studied contain two CPS enzymes and there has been a report suggesting that the pea *Pisum sativum* contains only one CPS (Doremus 1986). However, no conclusive genetic studies have yet been performed on any plant to confirm the number of CPS enzymes in these organisms. Further study of the CPS enzymes of plants and various protists is needed, since the organisms studied to date do not reflect the true diversity present within the eukaryotes, and prevent a firm dating of the duplication event(s). Of the eukaryotic microorganisms, only S. *cerevisiae* and N. *crassa* have been confirmed to have two CPS enzymes, and only *P. falciparum* and *B. bovis* have been confirmed to have one CPS (Davis 1986; Flores, O'Sullivan, and Stewart 1994; Chansiri and Bagnara 1995). Based on the present evidence it is concluded that that a gene duplication forming the two CPSs seen in eukaryotes probably occurred after the divergence of the apicomplexan protozoans.

Within the specialized eukaryotic CPS enzymes, three clusters were commonly seen comprising the arginine-specific enzymes, the urea-cycle-specific enzymes (CPSI, CPSIII), and the pyrimidine-specific enzymes. The branching order of these clusters relative to each other could not be fully resolved; however, the CPS enzymes used in arginine biosynthesis (Arg) and the urea cycle (CPSI, CPSIII) never clustered together (figs. 2–4), suggesting a possible separate evolutionary origin for the two. It is possible that more than one gene duplication event occurred to create the several types of CPS observed in the eukaryotes. With an enzyme like CPS being utilized for two metabolic pathways, one can see how advantageous it would be for an organism to have two CPS enzymes and, therefore, for a gene duplication event such as this to occur a few times in evolution.

Evolution of the Organization of CPS Genes: the Intervening Sequence Between the Amidotransferase and Synthetase Domains of Proteobacterial CPS is Variable

Comparison of the intervening sequences between the amidotransferase and synthetase domains of CPS in different organisms shows that this sequence has varied significantly over time, particularly within the Proteobacteria examined so far. Confirming the branching order for these lineages, one can see that multiple deletions and insertions must have occurred between *carA* and *carB*. For example, an insertion event likely led to the formation of an open reading frame between the amidotransferase and synthetase genes of ***P. aeruginosa,*** since the corresponding ***E. coli*** and ***P. stutzeri*** genes have a similar structure comprising a very small intervening sequence. The sequence between *carA* and *carB* in N. ***gonorrhoeae varies*** in size between isolates; however, this is more likely a reflection of this organism, which is noted for its heterogeneity (O'Rourke and Spratt 1994; Lawson, Billowes, and Dillon 1995).

Proposed Evolutionary History of Carbamoylphosphate Synthetase

We propose the following summary for the description of the evolution of the CPS genes, based on this phylogenetic analysis and the studies of others (such as Nyunoya and Lusty 1983; Nyunoya, Broglie, and Lusty 1985; Hong et al. 1994). First an amidotransferase gene became associated with a kinase gene which duplicated to form the synthetase domain, or the kinase gene duplicated and then associated with the amidotransferase gene, in either case resulting in genes encoding the first primeval heterodimeric CPS. Within the bacteria, the Proteobacteria and Gram-positive bacteria diverged and a gene duplication event led to the formation of two CPS enzymes in the Gram-positives. In the Proteobacteria, significant insertions and deletions occurred in the sequence between the genes encoding the heterodimeric CPS. Meanwhile, significant evolution of CPS was occurring in the Eukarya after the Archaea diverged. First there was a fusion of the amidotransferase and synthetase genes to form a monomeric CPS in the eukaryotes. Then, some time after the divergence of the apicomplexan protozoans, the first gene duplication event occurred to form the two CPS enzymes observed in fungi and animals. Whether one or two gene duplications led to the formation of the arginine-specific and urea-cycle-specific CPSs is as yet unclear, but each subsequently underwent significant change: the arginine-specific CPS genes of fungi redivided to encode a heterodimeric CPS enzyme once again, and the arginine-specific CPS of vertebrates obtained acetylglutamate-binding ability, to become CPSIII-like. Then in the lineage leading to human, rat, and bullfrog, glutamine-binding ability was lost and this CPSIII became the CPSI used to harvest ammonia for the urea cycle. During this evolution of the eukaryotes, the pyrimidine-specific CPS was relatively unchanged, although it did undergo a fusion with other enzymes involved in the pyrimidine biosynthetic pathway to become part of a multifunctional protein.

In conclusion, the evolution of CPS has involved gene duplications, gene fusions, redivision of previously fused genes, gene translocations, deletions and insertions in sequence surrounding the genes, and mutations within the genes, resulting in changes in function. Further investigations of this enzyme and its gene sequence(s) in other organisms should continue to prove interesting. In general, CPS genes seem amenable to phylogenetic analysis: they show a branching order consistent with other phylogenetic analyses, they reveal gene duplications which have occurred through clustering of these genes within the trees, and they have formed from an initial ancient duplication and so can be used to root the tree of life.

Acknowledgments

The authors wish to thank A. T. Abdelal and C.-D. Lu (Georgia State University, Atlanta, Ga.) for supplying CPS (synthetase) sequences from ***P. stutzeri*** and ***P. aeruginosa,*** J. R. Brown (Dalhousie University, Halifax, Nova Scotia) and G. Drouin (University of Ottawa, Ottawa, Ontario) for comments regarding this manuscript, and C. W. Sensen and others of the Sulfolobus Genome Project (R.L.C., M. A. Ragan, and W. F. Doolittle, principal investigators). This project was funded by the Canadian Bacterial Diseases Network, with the exception of the determination of the S. *solfataricus* sequence, which was funded by the Canadian Genome Analysis and Technology Program, the Canadian Institute for Advanced Research, and the National Research Council of Canada.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1993. Molphy (maximum likelihood inference of protein phylogeny). Version 2.1.2. Graduate University for Advanced Study, and Institute for Statistical Mathematics, Tokyo, Japan.

ANDERSON, P. M. 1980. Glutamine- and *N*-acetylglutamate-dependent carbamoyl phosphate synthetase in elasmobranchs. Science **208**:291–293.

BRINGEL, F. 1994. Isolation of Lactobacilli naturally deficient in carbamyl phosphate biosynthesis. XIVth International Arginine Symposium, September 12-14, 1994. Paris, France.

BROWN, J. R., and W. F. DOOLITTLE. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. USA 92: 2441-2445.

CHANSIRI, K., and A. S. BAGNARA. 1995. The structural gene for carbamoyl phosphate synthetase from the protozoan parasite *Babesia bovis.* Mol. Biochem. Parasitol. **74**:239–243.

CRETI, R., E. CECCARALLI, M. BOCCHETTA, A. M. SANANGELANTONI, 0. TIBONI, I? PALM, and P. CAMMARANO. 1994. Evolution of translational elongation factor (EF) sequences: reliability of global phylogenies inferred from EF-1 alpha (Tu) and EF-2(G) proteins. Proc. Natl. Acad. Sci. USA 91: 3255–3259.

DAVIDSON, J. N., K. C. CHEN, R. S. JAMISON, L. A. MUSMANNO, and C. B. KERN. 1993. The evolutionary history of the first three enzymes in pyrimidine biosynthesis. Bioessays **15:** 157-164.

DAVIS, R. H. 1986. Compartmental and regulatory mechanisms in the arginine pathways of *Neurospora crassa* and *Saccharomyces cerevisiae.* Microbiol. Rev. **50**:280–313.

DAVIS, R. H., J. L. RISTOW, and B. A. HANSON. 1980. Carbamoylphosphate synthetase A of *Neurospora crassa*. J. Bacteriol. **141**:144–155.

DOREMUS, H. D. 1986. Organization of the pathway of de novo pyrimidine nucleotide biosynthesis in Pea (*Pisum sativum* L. cv progress no. *9)* leaves. Arch. Biochem. Biophys. **250**: 112-l 19.

FELSENSTEIN, J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Department of Genetics, University of Washington, Seattle.

FLORES, M. V. C., W. J. O'SULLIVAN, and T. S. STEWART. 1994. Characterization of the carbamoyl phosphate synthetase gene from *Plasmodium falciparum.* Mol. Biochem. Parasitol. **68**:315–318.

FORTERRE, P., N. BENACHENHOU-LAHFA, F. CONFALONIERI, M. DUGUET, C. ELIE, and B. LABEDAN. 1993. The nature of the last universal ancestor and the root of the tree of life, still open questions. Biosystems 28: 15-32.

GESCHWILL, K., and L. LUMPER. 1989. Identification of cysteine residues in carbamoyl-phosphate synthase I with reactivity enhanced by *N*-acetyl-L-glutamate. Biochem. J. **260**:573–576.

GOGARTEN, J. I?, H. KILBAK, P. DITTRICH, L. TAIZ, E. J. BOWMAN, B. J. BOWMAN, M. F. MANOLSON, R. J. POOLE, T. DATE, and T. OSHIMA. 1989. Evolution of vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA **86**:6661–6665.

HELBING, C., and B. G. ATKINSON. 1994. Heat shock stabilizes T3-induced urea cycle enzyme gene expression in the liver of *Rana catesbeiana* tadpoles. J. Biol. Chem. 269: 11743-l 1750.

HILARIO, E., and J. F! GOGARTEN. 1993. Horizontal transfer of ATPase genes-the tree of life becomes a net of life. Biosystems *31:* 11 l-l 19.

HONG, J., W. L. SALO, C. J. LUSTY, and P. M. ANDERSON. 1994. Carbamyl phosphate synthetase III, an evolutionary intermediate in the transition between glutamine-dependent and ammonia-dependent carbamyl phosphate synthetases. J. Mol. Biol. **243**:131–140.

IWABE, N., K. KUMA, M. HASEGAWA, S. OSAWA, and T. MIYATA. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86**:9355–9359.

KILSTRUP, M., C.-D. Lu, A. T. ABDELAL, and J. NEUHARD. 1988. Nucleotide sequence of the *carA* gene and regulation of the *carAB* operon in *Salmonella typhimurium.* J. Biochem. **176**:421–429.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 11 l-120.

KWON, D.-H., C.-D. Lu, D. A. WALTHALL, T. M. BROWN, J. E. HOUGHTON, and A. T. ABDELAL. 1994. Structure and regulation of the *carAB* operon in *Pseudomonas aeruginosa* and *Pseudomonas stutzeri:* no untranslated region exists. J. Bacteriol. **176**:2532–2542.

LACROUTE, F., A. PIERARD, M. GRENSON, and J. M. WIAME. 1965. The biosynthesis of carbamoylphosphate in *Saccharomyces cerevisiae.* J. Gen. Microbiol. 40: 127-142.

LAKE, J. A., E. HENDERSON, M. OAKES, and M. W. CLARK. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci. USA **81**:3786–3790.

LAWSON, F. S., F. M. BILLOWES, and J. R. DILLON. 1995. Organization of carbamoyl-phosphate synthase genes in *Neisseria gonorrhoeae* includes a large, variable intergenic sequence which is also present in other *Neisseria* species. Microbiology **141**:1183–1191.

LEGRAIN, C., M. DEMAREZ, N. GLANSDORFF, and A. PIERARD. 1995. Ammonia-dependent synthesis and metabolic channeling of carbamoyl phosphate in the hyperthermophilic archaeon *Pyrococcus furiosus.* Microbiology 141: 1093-1099.

MARSHALL, M. 1976. Carbamoyl-phosphate synthetase I from frog liver. Pp. 133-142 *in S.* GRISOLIA, R. BAGUENA, and F. MAYOR, eds. The urea cycle. John Wiley and Sons, New York.

MIRAN, S. G., S. H. CHANG, and F. M. RAUSHEL. 1991. Role of four conserved histidine residues in the amidotransferase domain of carbamoyl phosphate synthetase. Biochemistry **30**:7901–7907.

NYUNOYA, H., K. E. BROGLIE, and C. J. LUSTY. 1985. The gene coding for carbamoylphosphate synthetase I was formed by fusion of an ancestral glutaminase gene and a synthetase gene. Proc. Natl. Acad. Sci. USA **82**:2244–2246.

NYUNOYA, H., and C. J. LUSTY. 1983. The *carB* gene of *Escherichia coli*: a duplicated gene coding for the large subunit of carbamoyl-phosphate synthetase. Proc. Natl. Acad. Sci. USA **80**:4629–4633.

O'ROURKE, M., and B. G. SPRATT. 1994. Further evidence for the non-clonal population structure of *Neisseria gonorrhoeae:* extensive genetic diversity within isolates of the same electrophoretic type. Microbiology 140: 1285-1290.

PAULUS, T. J., and R. L. SWITZER. 1979. Characterisation of pyrimidine-repressible and arginine-repressible carbamyl phosphate synthetases from *Bacillus subtilis.* J. Bacteriol. **137**:82–91.

PICARD, F. J., and J. R. DILLON. 1989. Cloning and organization of seven arginine biosynthesis genes from *Neisseria gonorrhoeae.* J. Bacteriol. **171**:1644–1651.

POST, L. E., D. J. POST, and F. M. RAUSHEL. 1990. Dissection of the functional domains of *Escherichia coli* carbamoyl phosphate syntbetase by site-directed mutagenesis. J. Biol. Chem. **265**:7742–7747.

REISER, J., V. GLUMOFF, U. OCHSNER, and A. FIECHTER. 1994. Molecular analysis of the *Trichosporin cutaneum* DSM 70698 *argA* gene and its use for DNA-mediated transformations. J. Bacteriol. **176**:3021–3032.

RUBINO, S. D., H. NYUNOYA, and C. J. LUSTY. 1986. Catalytic domains of carbamyl phosphate synthetase. Glutamine-hydrolyzing site of *Escherichia coli* carbamyl phosphate. J. Biol. Chem. 261: 11320-l 1327.

SCHOFIELD, J. I? 1993. Molecular studies on an ancient gene encoding for carbamoyl-phosphate synthetase. Clin. Sci. 84: 119-128.

SENSEN, C. W., R. L. CHARLEBOIS, R. K. SINGH, H.-P. KLENK, M. A. RAGAN, and W. F. DOOLITTLE. 1996. Sequencing the genome of *Sulfolobus solfataricus P2. In* F. J. DE BRUIJN, J. R. LUPSKI, and G. WEINSTOCK, eds. Bacterial genomes: physical structure and analysis. Chapman and Hall, New York (in press).

SIMMER, J. P., R. E. KELLY, A. G. RINKER JR., J. L. SCULLY, and D. R. EVANS. 1990. Mammalian carbamyl phosphate synthetase (CPS). cDNA sequence and evolution of the CPS domain of the syrian hamster multifunctional protein CAD. J. Biol. Chem. **265**:10395–10402.

VAN DEN HOFF, M. J. B., A. JONKER, J. J. BEINTEMA, and W. H. LAMERS. 1995. Evolutionary aspects of the carbamoylphosphate synthetase genes. J. Mol. Evol. **41**:813–832.

WERNER, M., A. FELLER, and A. PIÉRARD. 1985. Nucleotide sequence of the yeast gene *CPA1* encoding the small subunit of arginine-pathway carbamoylphosphate synthetase. Homology of deduced amino acid sequence to other glutamine amidotransferases. Eur. J. Biochem. **146**:37 l-38 1.

GARY J. OLSEN, reviewing editor

Accepted May 24, 1996