Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 9975–9979, December 1990
Biochemistry

Conserved enzymes mediate the early reactions of carotenoid biosynthesis in nonphotosynthetic and photosynthetic prokaryotes

(ADP-binding fold/carotenoid genes/*Erwinia herbicola*/*Rhodobacter capsulatus*/tomato fruit-ripening cDNA)

GREGORY A. ARMSTRONG*†‡§, MARIE ALBERTI†, AND JOHN E. HEARST*†

*Department of Chemistry, University of California and †Division of Chemical Biodynamics, Lawrence Berkeley Laboratory, Berkeley, CA 94720

Communicated by James Bonner, August 24, 1990

ABSTRACT **Carotenoids comprise one of the most widespread classes of pigments found in nature. The first reactions of C_{40} carotenoid biosynthesis proceed through common intermediates in all organisms, suggesting the evolutionary conservation of early enzymes from this pathway. We report here the nucleotide sequence of three genes from the carotenoid biosynthesis gene cluster of *Erwinia herbicola*, a nonphotosynthetic epiphytic bacterium, which encode homologs of the CrtB, CrtE, and CrtI proteins of *Rhodobacter capsulatus*, a purple nonsulfur photosynthetic bacterium. CrtB (prephytoene pyrophosphate synthase), CrtE (phytoene synthase), and CrtI (phytoene dehydrogenase) are required for the first three reactions specific to the carotenoid branch of general isoprenoid metabolism. The homologous proteins from *E. herbicola* and *R. capsulatus* show sequence identities of 41.7% for CrtI, 33.7% for CrtB, and 30.8% for CrtE. *E. herbicola* and *R. capsulatus* CrtI also display 27.2% and 27.9% sequence identity, respectively, with *R. capsulatus* CrtD (methoxyneurosporene dehydrogenase). All three dehydrogenases possess a hydrophobic N-terminal domain containing a putative ADP-binding $\beta\alpha\beta$ fold characteristic of enzymes known to bind FAD or NAD(P) cofactors. In addition, *E. herbicola* and *R. capsulatus* CrtB show 25.2% and 23.3% respective sequence identities with the protein product of pTOM5, a tomato cDNA of unknown function that is differentially expressed during fruit ripening. These data indicate the structural conservation of early carotenoid biosynthesis enzymes in evolutionarily diverse organisms.**

Carotenoids, a major class of natural pigments, serve a variety of biological functions including protection against photooxidative damage, an auxiliary light-harvesting function in photosynthesis, and the coloration of many birds, fish, insects, and marine invertebrates (for review, see ref. 1, p. 243). Carotenoid derivatives play crucial biological roles in humans and animals for nutrition (vitamin A), for the visual system (retinal), and as cellular growth regulators (retinoic acid) (1, p. 67), or in plants as hormones (abscisic acid) (2). Carotenoids are also of interest as natural colorants for food (1, p. 68), as possible anticancer agents (3), and as immune system enhancers (4).

All photosynthetic prokaryotes and eukaryotes, as well as certain fungi, yeasts, and nonphotosynthetic bacteria, synthesize carotenoids (1, pp. 39–43). The early reactions of general isoprenoid metabolism specific to carotenoid biosynthesis proceed through common intermediates to phytoene, the first C_{40} carotenoid (Fig. 1). Later variations in the biosynthetic pathway, such as the number of sequential dehydrogenations proceeding from phytoene through phytofluene, ζ-carotene, and neurosporene to lycopene, ring cyclizations, and insertion of oxygen-containing functional groups, generate the tremendous diversity of carotenoid species observed in nature (1). Roughly 600 chemically distinct carotenoids and their glycosides have thus far been identified (7).

It is not known whether the early reactions of carotenoid biosynthesis common to all organisms are mediated by enzymes of conserved structure. Although phytoene synthase from red pepper has been isolated (8), no purified prokaryotic enzyme involved in either phytoene synthesis or dehydrogenation has yet been described. We have chosen to study the evolutionary conservation of carotenoid biosynthesis enzymes by characterizing the corresponding genes.

Rhodobacter capsulatus, a purple nonsulfur photosynthetic bacterium, contains a cluster of eight *crt* genes devoted to carotenoid biosynthesis and is the only system from which *crt* genes have thus far been molecularly characterized (5, 9, 10). The products of *crtI*, *crtB*, and *crtE*, and a fourth gene unlinked to the *crt* gene cluster (*crtJ*) (11) are required for early biosynthetic reactions (Fig. 1). Biochemical and immunological data demonstrate that CrtI is the phytoene dehydrogenase (5, 12–14) and strongly suggest that CrtB and CrtE are the prephytoene pyrophosphate (PPPP) and phytoene synthases, respectively (nomenclature as in ref. 8). We have characterized carotenoid biosynthesis genes from a nonphotosynthetic epiphytic prokaryote, *Erwinia herbicola*, to investigate the evolutionary conservation of the enzymes catalyzing the earliest biosynthetic reactions. A preliminary report indicated that *E. herbicola* accumulates cyclic carotenoids such as β-carotene, β-cryptoxanthin, zeaxanthin and zeaxanthin glycosides (unpublished results quoted in ref. 15). A cluster of *E. herbicola* genes that direct the synthesis of yellow pigments in *Escherichia coli*, a normally unpigmented organism, were subsequently cloned (16). Initial characterization of the pigments suggested that they were polar β-carotene derivatives (17), a result supported by inhibitor studies (18). *R. capsulatus*, in contrast, synthesizes acyclic carotenoids derived from neurosporene. These data indicate that the *E. herbicola* and *R. capsulatus* carotenoid biosynthesis pathways diverge after phytoene dehydrogenation and before lycopene cyclization.

We report here the nucleotide sequences of three genes from the *E. herbicola crt* gene cluster¶, which encode products homologous to enzymes catalyzing the first three reactions committed to carotenoid biosynthesis in *R. capsulatus*.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: PPPP, prephytoene pyrophosphate; ORF, open reading frame; *Rc*-CrtE, *Rc*-CrtD, *Rc*-CrtI, and *Rc*-CrtB *Rhodobacter capsulatus* CrtE, CrtD, CrtI, and CrtB, respectively; *Eh*-CrtI, *Eh*-CrtE, and *Eh*-CrtB, *Erwinia herbicola* CrtI, CrtE, and CrtB, respectively.

‡Present address: Max-Planck-Institut für Züchtungsforschung, Abteilung Biochemie, D-5000 Cologne 30, Federal Republic of Germany.

§To whom reprint requests should be addressed.

¶The sequences reported in this paper have been deposited in the GenBank data base (accession nos. M38423 for *Erwinia herbicola crtI* and *crtB* and M38424 for *E. herbicola crtE*).

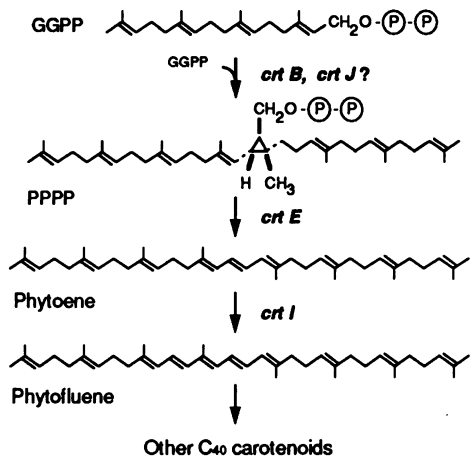

FIG. 1. Early reactions of carotenoid biosynthesis. The first reaction specific to carotenoid biosynthesis condenses two molecules of geranylgeranyl pyrophosphate (GGPP) to yield PPPP (1, pp. 46–47). The *crtB*, *crtE*, *crtI*, and *crtJ* gene products are required for the early steps of carotenoid biosynthesis in *R. capsulatus* (see ref. 5 for a description of the entire pathway). The role of the *crtJ* gene product has not been unambiguously determined. Phytoene and phytofluene occur in cis conformations in some systems (1, p. 48; 6).

Comparison of the deduced amino acid sequences of the *E. herbicola* and *R. capsulatus* gene products provides evidence that structurally similar enzymes mediate the earliest reactions of carotenoid biosynthesis in both nonphotosynthetic and photosynthetic prokaryotes.

MATERIALS AND METHODS

Plasmids, Cloning Techniques, and Nucleotide Sequence Determination. Plasmid pPL376 (16), carrying the *E. herbicola crt* gene cluster, was subcloned by ligating total digests of *Bam*HI-restricted or *Pst* I-restricted pPL376 DNA into the cloning vector pBR325 (19), digested with the appropriate restriction enzyme. Recombinant M13 phages for sequencing were generated either by shotgun cloning of restriction fragments directly from pPL376 or by recloning *E. herbicola* DNA inserts from the pBR325 recombinants into M13. DNA sequencing was performed by the dideoxy nucleotide chain-termination method (20) with either a commercial oligonucleotide primer (P-L Biochemicals) complementary to the M13 vector or with synthetic oligonucleotides (Applied Biosystems 381A DNA synthesizer) complementary to the insert DNA. Primers were extended by using the Klenow fragment of DNA polymerase I (BRL), and the extended fragments were labeled with [α^{32}P]dATP (400 Ci/mmol; 1 Ci = 37 GBq) (Amersham). All nucleic acid and enzymatic manipulations were done according to standard published procedures (21) or manufacturers' protocols.

Protein Sequence Comparison. Deduced amino acid sequences of open reading frames (ORFs) from the *E. herbicola crt* gene cluster were compared with the *crt* gene products from *R. capsulatus* as described (9). Three *E. herbicola* ORFs with sequence similarities to the *R. capsulatus* CrtI and CrtD, CrtB, and CrtE proteins, respectively, were identified with this procedure. Final sequence alignments and the insertion of gaps were performed manually by using the unitary-matrix alignment method, and sequence identities were calculated dividing by the shorter of the two sequences (22). *E. herbicola* and *R. capsulatus* protein sequences were compared against the National Biomedical Research Foundation (release 21.0, 6/89) and Swissprot (release 12.0 10/89) data bases using the FASTA program provided in the GCG sequence analysis software package (version 6.1, 8/89).

RESULTS

Nucleotide Sequences of *crtI*, *crtB*, and *crtE* Genes of *E. herbicola*. The *E. herbicola crt* gene cluster contains ORFs corresponding to the *crtI*, *crtB*, and *crtE* genes of *R. capsulatus* (see below) (9), which encode three early carotenoid biosynthesis enzymes (Fig. 1). Fig. 2 shows the nucleotide sequence of 3378 base pairs (bp) from the *E. herbicola* gene

A *crtI* --->

```
GAGATAAAGGATGAAAAAAACCGTTGTGATTGGCGCAGGCTTTGGTGGCCTGGCGCTGGCGATTCGCCTGCAGGCGGCAGGGATCCCAACCGTACTGCTGGAGCAGCGGGACAAGCCCGG  120
CGGTCGGGCCTACGTCTGGCATGACCAGGGCTTTACCTTTGACGCCGGGCCGACGGTGATCACCGATCCTACCGCGCTTGAGGCGCTGTTCACCCTGGCCGGCAGGCGCATGGAGGATTA  240
CGTCAGGCTGCTGCCGGTAAAACCCTTCTACCGACTCTGCTGGGAGTCCGGGAAGACCCTCGACTATGCTAACGACAGCGCCGAGCTTGAGGCGCAGATTACCCAGTTCAACCCCCGCGA  360
CGTCGAGGGCTACCGGCGCTTTCTGGCTTACTCCCAGGCGGTATTCCAGGAGGGATATTTGCGCCTCGGCAGCGTGCCGTTCCTCTCTTTTCGCGACATGCTGCGCGCCGGGCCGCAGCT  480
GCTTAAGCTCCAGGCGTGGCAGAGCGTCTACCAGTCGGTTTCGCGCTTTATTGAGGATGAGCATCTGCGGCAGGCCTTCTGGTTCCACTCCCTGCTGGTAGGCGGCAACCCCTTCACCAC  600
CTCGTCCATCTACACCCTGATCCACGCCCTTGAGCGGGAGTGGGGGGTCTGGTTCCCTGAGGGCGGCACCGGGGCGCTGGTGAACGGCATGGTGAAGCTGTTTACCGATCTGGGCGGGGA  720
GATCGAACTCAACGCCCGGGTCGAAGAGCTGGTGGTGGCCGATAACCGCGTAAGCCAGGTCCGGCTGGCGGATGGTCGGATCTTTGACACCGACGCCGTAGCCTCGAACGCTGACGTGGT  840
GAACACCTATAAAAAGCTGCTCGGCCACCATCCGGTGGGGCAGAAGCGGGCGGCAGCGCTGGAGCGCAAGAGCATGAGCAACTCGCTGTTTGTGCTCTACTTCGGCCTGAACCAGCCTCA  960
TTCCCAGCTGGCGCACCATACCATCTGTTTTGGTCCCCGCTACCGGGAGCTGATCGACGAGATCTTTACCGGCAGCGCGCTGGCGGATGACTTCTCGCTCTACCTGCACTCGCCCTGCGT 1080
GACCGATCCCTCGCTCGCGCCTCCCGGCTGCGCCAGCTTCTACGTGCTGGCCCCGGTGCCGCATCTTGGCAACGCGCCGCTGGACTGGGCGCAGGAGGGGCCGAAGCTGCGCGACCGCAT 1200
CTTTGACTACCTTGAAGAGCGCTATATGCCCGGCCTGCGTAGCCAGCTGGTGACCCAGCGGATCTTTACCCCGGCAGACTTCCACGACACGCTGGATGCGCATCTGGGATCGGCCTTCTC 1320
CATCGAGCCGCTGCTGACCCAAAGCGCCTGGTTCCGCCGCACAACCGCGACAGCGACATTGCCAACCTCTACCTGGTGGGCGCAGGTACTCACCCTGGGGCGGGCATTCCTGGCGTAGT 1440
```

crtB --->

```
GGCCTCGGCGAAAGCCACCGCCAGCCTGATGATTGAGGATCTGCAATGAGCCAACCGCCGCTGCTTGACCACGCCACGCAGACCATGGCCAACGGCTCGAAAAGTTTTGCCACCGCTGCG 1560
AAGCTGTTCGACCCGGCCACCCGCCGTAGCGTGCTGATGCTCTACACCTGGTGCCGCCACTGCGATGACGTCATTGACGACCAGACCCACGGCTTCGCCAGCGAGGCCGCGGCGGAGGAG 1680
GAGGCCACCCAGCGCCTGGCCCGGCTGCGCACGCTGACCCTGGCGGCGTTTGAAGGGGCCGAGATGCAGGATCCGGCCTTCGCTGCCTTTCAGGAGGTGGCGCTGACCCACGGTATTACG 1800
CCCCGCATGGCGCTCGATCACCTCGACGGCTTTGCGATGGACGTGGCTCAGACCCGCTATGTCACCTTTGAGGATACGCTGCGCTACTGCTATCACGTGGCGGGCGTGGTGGGTCTGATG 1920
ATGGCCAGGGTGATGGGCGTGCGGGATGAGCGGGTGCTGGATCGCGCCTGCGATCTGGGGCTGGCCTTCCAGCTGACGAATATCGCCCGGGATATTATTGACGATGCGGCTATTGACCGC 2040
TGCTATCTGCCCGCCGAGTGGCTGCAGGATGCCGGGCTGACCCCGGAGAACTATGCCGCGCGGGAGAATCGGGCCGCGCTGGCGCGGGTGGCGGAGCGGCTTATTGATGCCGCAGACCG 2160
TACTACATCTCCTCCCAGGCCGGGCTACACGATCTGCCGCCGCGCTGCGCCTGGGCGATCGCCACCGCCCGCAGCGTCTACCGGGAGATCGGTATTAAGGTAAAAGCGGCGGGAGGCAGC 2280
GCCTGGGATCGCCGCCAGCACACCAGCAAAGGTGAAAAAATTGCCATGCTGATGGCGGCACCGGGGCAGGTTATTCGGGCGAAGACGACGAGGGTGACGCCGCGTCCGGCCGGTCTTTGG 2400
CAGCGTCCGTTTAG 2415
```

B *crtE* --->

```
TGACTGATACTAAAAGACAATTCAGCGGGTAACCTTGCAATGGTGAGTGGCAGTAAAGCGGGCGTTTCGCCTCATCGCGAAATAGAAGTAATGAGACAATCCATTGACGATCACCTGGCT  120
GGCCTGTTACCTGAAACCGACAGCCAGGATATCGTCAGCCTTGCGATGCGTGAAGGCGTCATGGCACCCGGTAAACGGATCCGTCCGCTGCTGATGCTGCTGGCCGCCCGCGACCTCCGC  240
TACCAGGGCAGTATGCCTACGCTGCTCGATCTCGCCTGCGCCGTTGAACTGACCATACCGCGTCGCTGATGCTCGACGACATGCCCTGCATGGACAACGCCGAGCTGCGCCGCGGTCAG  360
CCCACTACCCACAAAAAATTTGGTGAGAGCGTGGCGATCCTTGCCTCCGTTGGGCTGCTCTCTAAAGCCTTTGGTCTGATCGCCGCCACCGGCGATCTGCCGGGGGAGAGGCGTGCCCAG  480
GCGGTCAACGAGCTCTCTACCGCCGTGGGCGTGCAGGGCCTGGTACTGGGGCAGTTTCGCGATCTTAACGATGCCGCCCTCGACCGTACCCCTGACGCTATCCTCAGCACCAACCACCTC  600
AAGACCGGCATTCTGTTCAGCGCGATGCTGCAGATCGTCGCCATTGCTTCCGCCTCGTCGCCGAGCACGCGAGAGACGCTGCACGCCTTCGCCCTCGACTTCGGCCAGGCGTTTCAACTG  720
CTGGACGATCTGCGTGACGATCACCCGGAAACCGGTAAAGATCGCAATAAGGACGCGGGAAAATCGACGCTGGTCAACCGGCTGGGCGCAGACGCGGCCCGGCAAAAGCTGCGCGAGCAT  840
ATTGATTCCGCCGACAAACACCTCACTTTTGCCTGTCCGCAGGGCGGCGCCATCCGACAGTTTATGCATCTGTGGTTTGGCCATCACCTTGCCGACTGGTCACCGGTCATGAAAATCGCC  960
TGA 963
```

FIG. 2. Nucleotide sequences of *E. herbicola crtI* and *crtB* genes (*A*) and *crtE* gene (*B*). Putative ribosome-binding sites (underlined) precede the proposed start codons (enlarged type). Stop codons mentioned in text are indicated in boldface type. Positions with respect to the first nucleotide shown are at right.

cluster containing these three genes. As in *R. capsulatus* (5, 9), *E. herbicola crtI* and *crtB* are adjacent and form a potential *crtIB* operon, whereas *crtE* is physically separated within the respective gene clusters (unpublished work). No other ORFs from the *E. herbicola* gene cluster display substantial amino acid sequence similarity to carotenoid biosynthetic enzymes of *R. capsulatus* (unpublished work). The organization of the entire *E. herbicola crt* gene cluster will be described elsewhere.

Fig. 2*A* shows the nucleotide sequence of a 2415-bp region containing the *crtI* and *crtB* genes of *E. herbicola*, with the proposed ATG start codon of *crtI* at bp 11, six nucleotides downstream from an in-frame stop codon. The TGA stop codon for *crtI* (bp 1487) overlaps the putative ATG start codon for *crtB* at bp 1486, a situation identical to that observed in *R. capsulatus* (9). The putative start codon of *E. herbicola crtB*, 12 nucleotides downstream from an in-frame stop codon, initiates an ORF extending to a TAG stop codon at bp 2413.

Fig. 2*B* shows the nucleotide sequence of a 963-bp region containing the *E. herbicola crtE* gene. The choice of an ATG at bp 40 as the *crtE* start codon maximizes the percent identity in the *E. herbicola* CrtE (*Eh*-CrtE) and *R. capsulatus* CrtE (*Rc*-CrtE) sequence comparison (Fig. 3). No alternative start codons exist between this ATG and an upstream in-frame stop codon at bp 1. A TGA stop codon for *crtE* is found at bp 961.

Conserved *E. herbicola* and *R. capsulatus* Carotenoid Dehydrogenases Share a Putative ADP-Binding $\beta\alpha\beta$ Fold. The deduced amino acid sequences of *E. herbicola* CrtI (*Eh*-CrtI), *R. capsulatus* CrtI (*Rc*-CrtI), and *R. capsulatus* CrtD (*Rc*-CrtD) have been aligned for comparison (Fig. 3*A*). *Rc*-CrtD (methoxyneurosporene dehydrogenase) performs a dehydrogenation distinct from that mediated by CrtI (phytoene dehydrogenase) and specific to the later stages of acyclic carotenoid biosynthesis in certain photosynthetic bacteria (for a summary of biosynthetic pathway, see ref. 5). The deduced protein sequences of the three carotenoid dehydrogenases demonstrate remarkable similarity. *Eh*-CrtI displays 41.7% sequence identity with *Rc*-CrtI, and many of the nonidentical residues represent conservative amino acid substitutions. *Eh*-CrtI and *Rc*-CrtD show 27.2% sequence identity, whereas the two *R. capsulatus* dehydrogenases, *Rc*-CrtI and *Rc*-CrtD, are 27.9% identical, with sequence similarity extending beyond the previously reported highly conserved N- and C-terminal regions (9).

Conserved hydrophobic residues at the N termini of the three dehydrogenases (Fig. 3*A*) shows strong sequence similarity to the fingerprint for an ADP-binding pocket common to enzymes containing FAD or NAD(P) cofactors (Fig. 4) (23). This pocket forms a $\beta\alpha\beta$ fold and is often found at the N termini of dinucleotide-binding domains. A search of the data bases revealed significant sequence similarity between the N termini of *Eh*-CrtI, *Rc*-CrtI, *Rc*-CrtD, and a variety of

A

```
                 ++++++++++++++++++++++++++++++
Eh-CrtI            *KKT******F****L*I**Q*A*IPTVLLEQR*K****AYVWHDQ*FT**A***VI*D*TA*EA*FTLA**RMEDY*R*L*VK***RLCWES*KTLDYAN  101
Rc-CrtI    MSKNTEGMGRAVVIGAGLGGLAAAMRLGAKGYKVTVVDRLDRPGGRGSSITKGGHRFDLGPTIVTVPDRLRELWADCGRDFDKDVSLVPMEPFYTIDFPDGEKYTAYG  108
Rc-CrtD         MRSETDV*****RM******IGAA*A*LR****EAG*A***KARAVPTP*GPA*T***VL*MRHV*DA*F*A**TRAEEHLT*I*LPRLARHFW***SSLDLFT  104

Eh-CrtI    *S*ELE*QITQFN*R****--Y*R*LAYSQ*VFQE**LR**SV*FLSFR*MLRAG*QLLK*Q*WQ***QSVSRFI----EDE*L***F*F***L***N**TTS*I*T*  203
Rc-CrtI    DDAKVKAEVARISPGDVEG--FRHFMWDAKARYEFGYENLGRKPMSKLWDLIKVLPTFGWLRADRSVYGHAKKMV----KDDHLRFALSFHPLFIGGDPFHVTSMYIL  210
Rc-CrtD    *T-EANI*AI*AFA**K*AAA**R*DHLTTGLW*AFHRSVIAA*KPD**RIAAATV*RPQ*WPALRPGLTMRDLLAHHF**PR*AQLFGRYATYV**R*GATPAVLS*  211

Eh-CrtI    IHA**REW**WFPE**TG*LVNG*V*LF**L***IE**AR*E*LV*A-*NRVSQV**A**RIFDTDA*A****VVN***K**GHHPVGQKRAAALERKSMSNSLFVLY  310
Rc-CrtI    VSQLEKKFGVHYAIGGVQAIADAMAKVITDQGGEMRLNTEVDEILVSRDGKATGIRLMDGTELPAQVVVSNADAGHTYKRLLRNRDRWRWTDEK*DK*RW*MG***W*  318
Rc-CrtD    IW*A*VQ-**WAIRE*MHGV*A*L*R*AEAK*VRFHYGAKAKR*--*RKE*RV*AVEIET*VSI*CGACIF*G*P*ALRDG**G--*AA*ASM**SPRPAP*LSAW**A  315

Eh-CrtI    **LNQ---PHSQLA***ICF****R*LIDE**TGSA**D*F***L*S*C------****SL**P*CAS****A***H**NAPL-**AQ*GP*LRDRIFDYL***YM**  408
Rc-CrtI    FGTKGTAKMWKDVGHHTVVVGPRYKEHVQDIFIKGELAEDMSLYVHRPS------VTDPTAAPKGDDTFYVLSPVPNLGFDNGVDWSVEAEKYKAKVLKVIEERLLPG  420
Rc-CrtD    **ATP---IGV*LA**N*FFTADP--*LEFGPIGA**MP*EPT**ICAQDREMQAP*PEIERFEIIMNGPAGHQ*F*-----------Q*EAQCR*RTFPMLAAMG*TF  408

Eh-CrtI    LRSQLVTQRI***AD*H*TLDAH**SA**I**LLT*********RDSDIAN*************I*G*VA*AKAT*SL**E*LQ                            492
Rc-CrtI    VAEKITEEVVFTPETFRDRYLSPLGAGFSLEPRILQSAWFRPHNASEEVDGLYLVGAGTHPGAGVPSVIGSGELVAQ-MIPDAPKPETPAAAAPKARTPRAKAAQ  524
Rc-CrtD    SPDPE*-RALT**ALLSR*FPGS***IYGGS*EGTL-*T**RPL*RTGLK****A*G*********MALT**THA*R--------------*LLAD*ISA**  494
```

B

```
pTOM5      MSVALLWVVSPCDVSNGTSFMESVREGNRFFDSSRHRNLVSNERINRGGGKQTNNGRKFSVRSAILATPSGERTMTSEQMVYDVVLRQAALVKRQLRSTNELEVKPDI  108

Eh-CrtB          *SQPPLLDHATQTMAN**K**AT**KLFD**T-*RSV*M**TW**HC**VI*DQTHGFASE***EEEATQ**ARLRTLTL*AFEGAEMQ*P****FQ*VALT  101
Rc-CrtB          MIAEADMEVCRELIRTGSYSF-HAASRVLPARVRDPALALYAFCRVADDEVDEVGAPR-DKAAAVLKLGDRLEDI----YAGRPRNAPSDRAFAAVVEEFEM   96
pTOM5      PIPGNLGLLSEAYDR*G*VCAEYAKT*-NLGTMLMTPER*RAIW*I*VW**RT*EL**GPN*SY-ITP**LDRWEN****V----FN***F-DML*G*LSDT*SN*PV  209

Eh-CrtB    HGIT*RMALDH*D***M*VAQTR*V*FE*TLR*CYH**GV**L**ARV*G**D------ERV*D********F*LT*****IID**AID*CY**AE*LQDA*LT*ENY  203
Rc-CrtB    PRELPEAL---LEGFAWDAEGRWYHTLSDVQAYSARVAAAVGAMMCVLMRVRN------PDALARACDLGLAMQMSNIARDVGEDARAGRLFLPTDWMVEEGIDPQAF  195
pTOM5      DIQPFRDM---I**MRM*LRKSR*KNFDELYL*CYY**GT**L*SVPI*GIAPESKATTESVYNA*LA**I*N*LT**L********R***VY**Q*ELAQA*LSDEDI  314

Eh-CrtB    A*RENRAA-LA**A***IDA*EPY*ISSQA*LHD**PR*AWA*AT*RSV*RE**IK*KA*GGSAWD**Q**S**E*IA----ML*A*PGQVIRAKTT**TPR*AGLWQ  306
Rc-CrtB    LADPQPTKGIRRVTERLLNRADRLYWRAATGVRLLPFDCRPGIMAAGKIYAAIGAEVAKAKYDNITRRAHTTKGRKLWLVANSAMSATATSMLPLSPRVHAKPEPEVA  303
pTOM5      F*GRVTD*-W*IFMKKQIH**RKFFDE*EK**TE*SSAS*FPVW*SLVL*RK*LD*IEAND*N*F*K**YVS*SKQVDCITYCICKISCA*YKNA*LQR           412

Eh-CrtB    RP*                                                                                                        309
Rc-CrtB    HLVDAAAHRNLHPERSEVLISALMALKARDRGLAMD                                                                       339
```

C

```
Eh-CrtE    MVSGSKAGV*PHRE**VMRQSIDDHLAG*L**TDSQDIVS**MRE--**MAP*K****LLML*AARD*RYQGSM*TLL*L*C-*V**T*T***ML**M*CM***EL**  105
Rc-CrtE            MSLDKRIESA------LVKALSPEALGESPPLLAAALPYGVFPGGARIRPTI--LVSVALACGDDCPAVTDAAAVALELMHCASLVHDDLPAFDNADIRR   92

Eh-CrtE    *Q*TT**KFG*SV*I**SVG**SKA*GLI*AT*DLPGE*RAQAVNE*STAV*VQ-*LVL**FRDLNDAALDRTPDAILSTNH***G**F*AM*QI***ASASSPST*E  212
Rc-CrtE    GKPSLHKAYNEPLAVLAGDSLLIRGFEVLADVGAVNPDRALKLISKLGQLSGARGGICAGQAWESESKVD------LAAYHQAKTGALFIAATQMGAIAAGY---EAE  191

Eh-CrtE    TLHAFALDF*Q***LL***R*DHP---ET**DRNK*AG--KSTL*NRL*ADA***QK*REHIDS*DKHLTFA**Q*G*IRQF*HLWFG*HLA*WSPVMKIA          307
Rc-CrtE    PWFDLGMRIGSAFQIADDLKDALMSAEAMGKPAGQDIANERPNAVKTMGIEGARRKHLQDVLAGAIASIPS-CP-GEAKLAQMVQLYAHKIMDIPASAERG          289
```

FIG. 3. Amino acid sequence alignments of early carotenoid biosynthesis enzymes from *E. herbicola* (Eh) and *R. capsulatus* (Rc). –, Gaps; *, residues identical to the Rc-CrtI sequence; amino acid positions are numbered at right. (*A*) Sequence alignments of *Eh*-CrtI, *Rc*-CrtI, and *Rc*-CrtD (National Biomedical Research Foundation accession nos. A33120, S04402, and S04406). +, Highly conserved hydrophobic region containing putative ADP-binding fold (see Fig. 4). We originally proposed the GTG codon encoding Val-34 of *Rc*-CrtI to start the *crtI* coding region (9). We propose another start codon for *crtI* based on the extended N-terminal sequence similarity of *Eh*-CrtI and *Rc*-CrtD with *Rc*-CrtI. Because the ATG codons encoding Met-1 and Met-8 of *Rc*-CrtI are both preceded by possible ribosome-binding sites (10) and lacking additional identities in the sequence alignments with the other dehydrogenases, we show here the ATG encoding Met-1 as the tentative *R. capsulatus crtI* start codon. A similar conclusion has been independently reached by other researchers subsequent to our publication of *R. capsulatus crtI* sequence (14). (*B*) Sequence alignments of *Eh*-CrtB and *Rc*-CrtB (National Biomedical Research Foundation accession nos. B33120 and S04403) with tomato pTOM5 protein (Swissprot accession no. Pto5$Lyces). (*C*) Sequence alignments of *Eh*-CrtE and *Rc*-CrtE (National Biomedical Research Foundation accession nos. C33120 and S04407).

Fingerprint		ββββββ b s s G G	αααααααααααααα G s s	ββββββββ s s a	
Eh-CrtI	3	KTVVIGAGFGGLALAIRLQAAGIP-TVLLE			31
Rc-CrtI	10	RAVVIGAGLGGLAAAMRLGAKGYK-VTVVD			38
Rc-CrtD	6	DVVVIGARMGGLAAAIGAAAAGLR-VTVVE			34
Pf-Hh	4	QVAIIGAGPSGLLLGQLLHKAGID-NVILE			32
Hu-Gr	22	DYLVIGGGSGGLASARRAAELGAR-AAVVE			50
Ec-Sdh	9	DAVVIGAGGAGIAARALQISQSGQT-CALLS			37
Ec-Ddh	7	QVVVLGAGPAGYSAAFRCADLGLE-TVIVE			35
Hu-Ddh	43	DVTVIGSGPGGYVAAIKAAQLGFK-TVCIE			71
Ec-Ao	10	DVLIIGSGAAGLSLALRLADQHQV-IVLSK			38
Ec-Gs	148	KVAIIGAGPAGLACADVLTRNGVK-AVCFD			176
Ps-Tm	40	RVAIVGAGISGLVAATELLRAGVKDVVLYE			69
Rm-FixC	7	DAIVVGAGMSGNAAAYAMASRGLK-VLQLE			35
Consenus		IGAG G A G			

FIG. 4. Sequence similarities of the conserved N-terminal regions of the carotenoid dehydrogenases to other proteins containing an ADP-binding $\beta\alpha\beta$ fold for FAD or NAD(P) cofactors. The ADP-binding fold fingerprint is shown (residues are G, glycine; b, basic or hydrophilic; s, small or hydrophobic; a, acidic) (23). Positions of each sequence within the respective proteins are indicated; dashes show a gap. The consensus indicates that residues are conserved in at least 9 of 12 sequences. Arg-13 of *Rc*-CrtD (underlined) may represent the site of a point mutation (see text). Protein sequences, sources, and National Biomedical Research Foundation accession nos. are as follows: Pf-Hh, *p*-hydroxybenzoate hydroxylase of *Pseudomonas fluorescens* (A00507); Hu-Gr, glutathione reductase of human (A00404); Ec-Sdh, succinate dehydrogenase of *E. coli* (C28836); Ec-Ddh, dihydrolipoamide dehydrogenase of *E. coli* (A00405); Hu-Ddh, dihydrolipoamide dehydrogenase of human (B28448 and A28961; the latter sequence was identical to that of pig liver enzyme, accession no. A28448); Ec-Ao, L-aspartate oxidase of *E. coli* (S01132); Ec-Gs, small subunit of glutamate synthase (NADPH) of *E. coli* (B29617); Ps-Tm, tryptophan 2-monooxygenase of *Pseudomonas syringae* (A25493); Rm-FixC, FixC protein of *Rhizobium meliloti* (C26952). All of the above proteins, except for FixC, are known to bind FAD and/or NAD(P) cofactors.

other eukaryotic and prokaryotic proteins that bind FAD or NAD(P) cofactors (Fig. 4). A second highly conserved region of unknown function found in all three dehydrogenases (*Eh*-CrtI residues 444–474, *Rc*-CrtI residues 456–486, *Rc*-CrtD residues 444–472) (Fig. 3*A*) (9) shows no obvious relationship to other sequences present in the data bases. The predicted molecular mass of *Eh*-CrtI is 54 kDa, compared with 58 kDa for *Rc*-CrtI (10). Immunological studies have reported apparent molecular masses of 60–64 kDa for *Rc*-CrtI (13, 14).

The Conserved CrtB Proteins of *E. herbicola* and *R. Capsulatus* Show Homology to a Tomato Protein Associated with Fruit Ripening. *E. herbicola* CrtB (*Eh*-CrtB) shows 33.7% overall sequence identity with *R. capsulatus* CrtB (*Rc*-CrtB), including an exact alignment between the N termini of the two PPPP synthases (Fig. 3*B*). The most highly conserved feature in the two sequences is a 17-residue stretch with 14 identities (*Eh*-CrtB residues 160–176 and *Rc*-CrtB residues 152–168). The predicted molecular mass of *Eh*-CrtB is 34 kDa, as compared with 37 kDa for *Rc*-CrtB (9). *Eh*-CrtB and *Rc*-CrtB also show 25.2% and 23.3% respective sequence identities to the protein product of pTOM5 (Fig. 3*B*), a tomato cDNA clone differentially expressed during fruit ripening (24). No function has been ascribed to pTOM5, which encodes a polypeptide with an observed molecular mass of 48 kDa. The regions of greatest sequence similarity between *Eh*-CrtB and *Rc*-CrtB are largely conserved in the pTOM5 protein.

Comparison of Conserved CrtE Proteins of *E. herbicola* and *R. capsulatus*. *Eh*-CrtE (phytoene synthase) shows 30.8% sequence identity with *Rc*-CrtE, with exact alignment of the C termini (Fig. 3*C*). Clusters of identical amino acids are found in the central region of the two proteins. *Eh*-CrtE has a predicted molecular mass of 33 kDa compared with 30 kDa for *Rc*-CrtE (9). Neither protein showed similarity to other sequences in the data bases.

DISCUSSION

Despite the enormous structural diversity of carotenoids synthesized in nature, the early reactions of carotenoid biosynthesis (Fig. 1) proceed through common intermediates (1, pp. 46–48). *R. capsulatus* synthesizes acyclic carotenoids derived from neurosporene (5). *E. coli* strains carrying the *E. herbicola crt* genes accumulate derivatives of β-carotene (15, 17, 18), a carotenoid formed by dehydrogenation of neurosporene to lycopene and cyclization of both ends of the latter carotenoid (1, p. 56). Although the later stages of carotenoid biosynthesis diverge in *R. capsulatus* and *E. herbicola*, both organisms might be expected to possess similar enzymes for synthesis and dehydrogenation of phytoene, the precursor of all other C_{40} carotenoids.

The nucleotide sequence of the *E. herbicola crt* gene cluster supports this hypothesis by revealing genes analogous to *crtI*, *crtB*, and *crtE* from *R. capsulatus*, encoding three early carotenoid biosynthesis enzymes. For all amino acid alignments of Fig. 3, percentage of identical residues and normalized alignment scores (unpublished work) indicate that the protein sequences are homologous and, hence, evolutionarily related (22). None of the other five *R. capsulatus crt* gene products characterized to date (9), *Rc*-CrtD excepted, show significant similarity to any ORFs from the *E. herbicola* gene cluster. The similarity between *Rc*-CrtD and *Eh*-CrtI (Fig. 3*A*) is not surprising, as we have previously observed sequence similarities between *Rc*-CrtD and *Rc*-CrtI (9). The absence of further sequence similarities between carotenoid biosynthetic enzymes in *E. herbicola* and *R. capsulatus* presumably reflects divergence in the later portions of the two biosynthetic pathways.

Of the three early enzymes of carotenoid biosynthesis, CrtI, the phytoene dehydrogenase, exhibits the greatest sequence conservation between nonphotosynthetic and photosynthetic prokaryotes. *Eh*-CrtI and *Rc*-CrtI are 41.7% identical and also show many conservative amino acid substitutions. Immunological data from antibodies against an *Rc*-CrtI fusion protein foreshadowed this result, suggesting some structural conservation of CrtI among all photosynthetic organisms (13). Although this conservation remains to be demonstrated, our data clearly indicate a close relationship between the CrtI proteins from nonphotosynthetic and photosynthetic prokaryotes.

In light of the likely evolutionary relationship between CrtI and CrtD, it is intriguing to speculate that *Rc*-CrtD, which performs a specialized dehydrogenation, may have arisen by gene duplication from an ancestral protein involved in the more general reaction of phytoene dehydrogenation. The absence of a second putative dehydrogenase in the *E. herbicola crt* gene cluster when the criterion of sequence similarities between *Eh*-CrtI, *Rc*-CrtI, and *Rc*-CrtD (Fig. 3*A*) (unpublished data) is used argues that this organism possesses one enzyme for the four sequential dehydrogenations that convert phytoene to lycopene.

Putative ADP-binding pockets in the conserved N-terminal regions of all three carotenoid dehydrogenases suggest not only that these enzymes bind an FAD or NAD(P) cofactor but that this requirement has been evolutionarily preserved in both photosynthetic and nonphotosynthetic bacteria. No direct evidence is available on the cofactor requirements for phytoene dehydrogenation in either *R. capsulatus* or *E. herbicola*. Recent data argue, however, against a direct requirement for either FAD or NAD(P) cofactors for phytoene dehydrogenation in daffodil chromoplasts (6). Phytoene dehydrogenase itself remains to be purified or biochemically characterized.

The *Rc*-CrtD sequence was deduced from the nucleotide sequence of the *crtD223* allele (9), which contains an undetermined point mutation resulting in the accumulation of

yellow carotenoids (25). Interestingly, Arg-13 of *Rc*-CrtD sits in the heart of the normally hydrophobic putative ADP-binding fold (Fig. 4). A glycine occupies this position in other proteins that bind FAD or NAD(P) cofactors and forms a part of the strictly conserved core fingerprint for ADP-binding folds (23). This invariant glycine also closely approaches the pyrophosphate moiety of the dinucleotide, as determined by comparing the x-ray crystal structures of two FAD-binding domains (26). These authors proposed that the substitution of a residue other than glycine would sterically interfere with pyrophosphate binding. Arg-13 of *Rc*-CrtD is encoded by AGA, a rare codon in *R. capsulatus* genes (9), and could have been derived from a GGA triplet encoding glycine by a single-bp mutation in the first position of the codon. We speculate that the unusual occurrence of Arg-13 in a critical position within the putative *Rc*-CrtD ADP-binding fold could disrupt cofactor binding and may constitute the molecular basis of the *crtD223* mutant phenotype (5, 25).

Neither CrtB (PPPP synthase) nor CrtE (phytoene synthase) have been purified or otherwise biochemically characterized in *R. capsulatus*, although we have shown a *crtB* mutant to accumulate geranylgeranyl pyrophosphate *in vitro*, whereas a *crtE* mutant accumulates PPPP, leading to the proposal of specific enzymatic functions for the gene products (5). In contrast to *R. capsulatus*, a bifunctional enzyme combining both the PPPP- and phytoene-synthesizing activities in a single 47.5-kDa polypeptide has been purified from red pepper chromoplasts (8). The sequence similarity of the predicted tomato pTOM5 protein with *Eh*-CrtB and *Rc*-CrtB, although not with *Eh*-CrtE and *Rc*-CrtE, suggests the possible involvement of pTOM5 in carotenoid biosynthesis. pTOM5 is differentially expressed during tomato fruit ripening and encodes a 48-kDa polypeptide (24), in close agreement with the value reported for the phytoene synthase from red pepper fruits (8). Massive lycopene accumulation in chromoplasts endows ripening cultivated tomatoes with their characteristic reddish-orange color (1, p. 65). Expression of pTOM5 during tomato fruit ripening could stimulate the activity of the early carotenoid biosynthetic pathway. Hence, we propose that pTOM5 may encode the tomato homolog of the bifunctional red pepper phytoene synthase and, thus, represents the cDNA product of a higher plant carotenoid biosynthesis gene.

Erwinia and *Rhodobacter* branched evolutionarily from a common prokaryotic ancestor more than one billion years ago (27). Our results indicate that the amino acid sequences of the early carotenoid biosynthesis enzymes have been conserved during this period. Subsequent lateral gene transfer, facilitated by the clustering of most or all *crt* genes in photosynthetic bacteria such as *R. capsulatus* (9, 25) and in nonphotosynthetic bacteria such as *E. herbicola* (16) could provide an alternative explanation for the sequence conservation.

Structural conservation between early carotenoid biosynthesis enzymes from nonphotosynthetic and photosynthetic prokaryotes, and, perhaps higher plants, suggests intriguing possibilities for the engineering of hybrid biosynthetic pathways. Further work will reveal the degree of structural and functional conservation between prokaryotic and eukaryotic carotenoid biosynthesis enzymes.

Note added in proof. The nucleotide sequence of the *al-1* gene of *Neurospora crassa*, encoding phytoene dehydrogenase has recently been determined (28). Based on the deduced amino acid sequence, the fungal phytoene dehydrogenase also possesses an N-terminal ADP-binding fold (29).

We thank R. Tuveson for providing plasmid pPL376; M. Beikman for construction of pPL376 subclones; D. Koh for oligonucleotide synthesis; and P. Beyer, D. Burke-Agüero, D. Cook, C. Maulbecker, D. O'Brien, and A. Schmidt for advice and critical reading of this manuscript. This material is based upon work supported by a National Science Foundation Graduate Fellowship (G.A.A.), by National Institutes of Health Grant GM 30786 (J.E.H.), and by the Office of Basic Energy Sciences, Biological Energy Division, Department of Energy, under Contract DE-ACO30-76SF00098 (J.E.H.).

1. Britton, G. (1983) in *The Biochemistry of Natural Pigments* (Cambridge Univ. Press, Cambridge, U.K.).
2. Walton, D. C. (1988) in *Plant Hormones and Their Role in Plant Growth and Development*, eds. Davies, P. J. (Kluwer, Dordrecht, Netherlands), pp. 113–131.
3. Krinsky, N. I. (1989) *J Nutr.* **119**, 123–126.
4. Bendich, A. (1989) *J. Nutr.* **119**, 112–115.
5. Armstrong, G. A., Schmidt, A., Sandmann, G. & Hearst, J. E. (1990) *J. Biol. Chem.* **265**, 8329–8338.
6. Beyer, P., Mayer, M. & Kleinig, H. (1989) *Eur. J. Biochem.* **184**, 141–150.
7. Straub, O. (1987) in *Key to Carotenoids*, ed. Pfander, H. (Birkhauser, Basel), p. 296.
8. Dogbo, O., Laferriere, A., d'Harlingue, A. & Camara, B. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 7054–7058.
9. Armstrong, G. A., Alberti, M., Leach, F. & Hearst, J. E. (1989) *Mol. Gen. Genet.* **216**, 254–268.
10. Armstrong, G. A., Alberti, M., Leach, F. & Hearst, J. E. (1990) in *Molecular Biology of Membrane-Bound Complexes in Phototrophic Bacteria*, eds. Drews, G. & Dawes, E. (Plenum, New York), pp. 39–46.
11. Zsebo, K. M. & Hearst, J. E. (1984) *Cell* **37**, 937–947.
12. Giuliano, G., Pollock, D. & Scolnik, P. A. (1986) *J. Biol. Chem.* **261**, 12925–12929.
13. Schmidt, A., Sandmann, G., Armstrong, G. A., Hearst, J. E. & Böger, P. (1989) *Eur. J. Biochem.* **184**, 375–378.
14. Bartley, G. E. & Scolnik, P. A. (1989) *J. Biol. Chem.* **264**, 13109–13113.
15. Starr, M. P. (1981) in *The Prokaryotes: A Handbook on Habitats, Isolation, and Identification of Bacteria*, eds. Starr, M. P., Stolp, H., Trüper, H. G., Balows, A. & Schlegel, H. G. (Springer, New York), p. 753.
16. Perry, K. L., Simonitch, T. A., Harrison-Lavoie, K. J. & Liu, S. T. (1986) *J. Bacteriol.* **168**, 607–612.
17. Tuveson, R. W., Larson, R. A. & Kagan, J. (1988) *J. Bacteriol.* **170**, 4675–4680.
18. Armstrong, G. A. (1989) Ph.D. thesis (Univ. of California, Berkeley).
19. Bolivar, F. (1978) *Gene* **4**, 121–136.
20. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
21. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) in *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
22. Doolittle, R. F. (1986) in *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences* (University Science Books, Mill Valley, CA), pp. 3–60.
23. Wierenga, R. K., Terpstra, P. & Hol, W. G. J. (1986) *J. Mol. Biol.* **187**, 101–107.
24. Ray, J., Bird, C., Maunders, M., Grierson, D. & Schuch, W. (1987) *Nucleic Acids Res.* **15**, 10587.
25. Marrs, B. (1981) *J. Bacteriol.* **146**, 1003–1012.
26. Wierenga, R. K., Drenth, J. & Schulz, G. E. (1983) *J. Mol. Biol.* **167**, 725–739.
27. Wilson, A. C., Ochman, H. & Prager, E. M. (1987) *Trends Genet.* **3**, 241–247.
28. Schmidhauser, T. J., Lauter, F. R., Russo, V.E.A. & Yanofsky, C. (1990) *Mol. Cell. Biol.* **10**, 5064–5070.
29. Bartley, G. E., Schmidhauser T. J., Yanofsky, C. & Scolnik, P. A. (1990) *J. Biol. Chem.* **265**, 16020–16024.