Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 6614–6618, September 1990
Genetics

Genes for tryptophan biosynthesis in the archaebacterium *Haloferax volcanii*

(*Halobacterium*/genetic mapping/indole-3-glycerol-phosphate synthase/tryptophan synthase/evolution)

WAN L. LAM*, ANNALEE COHEN, DINA TSOULUHAS, AND W. FORD DOOLITTLE

Department of Biochemistry, Dalhousie University, Halifax, NS, B3H 4H7, Canada

Communicated by Charles Yanofsky, June 13, 1990

ABSTRACT **Recent technical advances permit direct genetic approaches for isolating genes and mapping auxotrophic mutations in the halophilic archaebacterium (Archaea) *Haloferax volcanii*. Twenty-nine mutations in tryptophan biosynthesis mapped to two separate chromosomal locations. DNA sequencing of one gene cluster shows a unique gene order (*trpCBA*) and unusual potential secondary structures in the 5′-flanking region.**

Inquiries into basic features of archaebacterial gene structure and function promise to provide an understanding of the earliest stages of cellular evolution (1, 2). To date, such inquiries have been limited by the lack of methods for genetic manipulation of archaebacteria [Archaea (3)] and thus have been restricted largely to indirect techniques for cloning archaebacterial genes into eubacterial hosts.

In this laboratory, we have begun to develop more direct approaches, allowing us to do archaebacterial genetics in archaebacteria. We can now efficiently transfect and transform spheroplasts of *Halobacterium halobium* and *Haloferax volcanii* with halobacterial phage, plasmid, and chromosomal DNAs (4, 5). We also have prepared shuttle-vectors selectable via drug resistance in *Escherichia coli* and *H. volcanii* (5) and are assembling a fine-scale restriction map of the *H. volcanii* genome by a "bottom up" approach (6). Here we apply several of these techniques to a description of some of the genes for tryptophan (*trp*) biosynthesis in *H. volcanii*.

Tryptophan biosynthetic genes were chosen because these and their products already make up a large and extensively characterized set of homologs. Crawford's very recent review (7) lists between 10 and 25 completed gene or protein sequences for each of the seven component polypeptides (E, G, D, C, F, B, and A) catalyzing the five steps of the *trp* pathway, with purple bacteria, Gram-positive bacteria, spirochetes, and fungi represented. Among archaebacteria, the *Methanococcus voltae trpB*, *trpA*, and part of *trpF* coding regions have been sequenced by Sibold and Henriquet (8), using a DNA fragment cloned by virtue of its ability to complement *E. coli* auxotrophs. With *H. volcanii*, direct genetic methods allow us to identify here two unlinked clusters of *trp* genes. Sequencing of the *trpB* region confirms our functional assignment and reveals an organization of genes within this cluster unlike that in eukaryotes, eubacteria, or *Methanococcus*.†

MATERIALS AND METHODS

Materials. Nucleotides and enzymes were purchased from Boehringer Mannheim and Pharmacia. Mevinolin was provided by A. Alberts (Merck Sharp & Dohme). Ethyl methanesulfonate and polyethylene glycol 600 (PEG 600) came from Sigma. PEG 600 was further purified as described (5).

Cell Strains and Plasmids. *H. volcanii* strain WFD11 (4) was produced by curing the parental strain DS2 of its endogenous 6.4-kilobase (kb) plasmid pHV2. Plasmid pHV51, a variant of pHV2, and the shuttle vector pWL102, which confers mevinolin resistance in *H. volcanii* and ampicillin resistance in *E. coli*, have been described (5). Cell strains were maintained on minimal or rich medium (4, 5). When necessary, minimal medium was supplemented with anthranilic acid (25 mg/liter), indole (5 mg/liter), or tryptophan (50 mg/liter) with or without mevinolin (10 μM).

Transformation of *H. volcanii*. Procedures were carried out at room temperature. Spheroplasts were transformed with plasmid DNAs using the PEG-mediated method described before (5). A multiplex system was devised for cosmid transformation. Three hundred microliters of 0.5 M EDTA (pH 8.0) was added to 2.4 ml of frozen spheroplasts of each Trp^- auxotroph. Aliquots (80 μl) of the mixtures were dispensed into 27 microtiter wells containing about 2 μg of DNA dissolved in 0.1 M EDTA. Pools of cosmid DNAs, constructed so that each of 154 cosmids was present in 2 of 25 different pools, were used. Two control transformations—no DNA and wild-type *H. volcanii* DNA—were included. Ninety microliters of a PEG solution [60% (vol/vol) PEG 600/0.4 M NaCl/10 mM KCl/6% sucrose/50 mM Tris·HCl, pH 7.5] was added and mixed by gentle pipetting with a multichannel pipettor. Transformation was terminated by adding 90 μl of regeneration solution (5). One-tenth volume of each transformation mixture was plated in top agar onto minimal agar plates. The small amount of PEG 600 and EDTA carried over does not noticeably hinder the regeneration of spheroplasts. Transformants appear after 7–10 days.

DNA Sequence and Computer Analysis. DNA fragments were subcloned into M13 vectors. Ordered sets of deletions were generated by using Exo nuclease III (9). Both strands were sequenced by the dideoxy chain-termination method (10) using the Taquenase system (United States Biochemical). Deoxyguanosine triphosphate was replaced by its 7-deaza analog in all sequencing reactions. Sequencing reaction products were resolved on polyacrylamide gels containing either 42% urea or 85% formamide as denaturant.

Deduced amino acid sequences were aligned with MULTALIN software (11). Sequences flanking the coding regions of the *trp* genes were analyzed for secondary structures by using MicroGenie programs from Beckman (12).

RESULTS AND DISCUSSION

Isolation and Identification of Mutants. We used a procedure modified from that described by Mevarech and Wercz-

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.

Abbreviation: ORF, open reading frame.

*To whom reprint requests should be addressed.

†This sequence has been deposited in the EMBL/GenBank data base (accession no. M36177).

Table 1. Summary of auxotrophic strains, their nutritional requirement, and the plasmids or cosmids that transform them to prototrophy

WFD strain	Allele number	Use	Cosmid(s) identified
35	*trp-101*	T	452
36	*trp-102*	T	452
37	*trp-103*	A, I	A159
39	*trp-105*	T	452
41	*trp-106*	T	452
131*	*trp-107*	T	452
135	*trp-108*	I	452
163	*trp-109*	I	A159
167	*trp-110*	I	A159
178	*trp-111*	I	A159
186	*trp-112*	I	A159
192	*trp-113*	I	A159
204	*trp-114*	I	A159
207	*trp-115*	I	A159
216	*trp-116*	I	A159
240	*trp-117*	I	A159
246	*trp-118*	I	488, A159
258	*trp-119*	I	A159
261*	*trp-120*	I	488, A159
284	*trp-121*	I	A159
287*	*trp-122*	T	452
292	*trp-123*	I	A159
300*	*trp-124*	I	488, A159
347	*trp-125*	T	452
369	*trp-131*	T	452
531*	*trp-127*	I	488, A159
543	*trp-128*	T	452
564	*trp-129*	I	A159
567	*trp-130*	I	A159

The indole-utilizing WFD135 and all the tryptophan-requiring (T) mutants were complemented by pDWT1, while the rest of the mutants that grow on indole (I) or anthranilic acid (A) were rescued by pT240-1, pT186-4, and pT1323-15.
*Strains initially transformed with cosmid pools.

berger (13) to produce auxotrophic mutants of *H. volcanii* WFD11 by treatment with ethyl methanesulfonate. Specific nutritional requirements were then identified by streaking on plates supplemented with various combinations of amino acids and nucleotides. Twenty-nine tryptophan-utilizing mutants were identified out of >400 auxotrophs isolated in three different mutant hunts. These Trp^- mutants were further categorized by their ability to grow on medium containing anthranilic acid, indole, or tryptophan (Table 1).

Cloning of *trp* Genes. We have cloned genes of tryptophan biosynthesis by preparing shotgun libraries of wild-type DNA in vectors derived from the endogenous *H. volcanii* plasmid pHV2 and transformation of Trp^- auxotrophs to prototrophy. Initially, two auxotrophs that would grow only on tryptophan, WFD35 and WFD36, were transformed with *Eco*RI-digested wild-type DNA cloned into compatible sites in the ISH51 insertion sequence of pHV51. [pHV51 is a fortuitously isolated variant of pHV2 containing a copy of the insertion sequence ISH51 (5).] Plasmid DNAs isolated from independent Trp^+ transformants (of either recipient) all contained a common 8-kb insert. In fact, all nine tryptophan-requiring Trp^- mutants could be transformed to prototrophy (6 to 8 thousand times more efficiently than with linear wild-type DNA) by either of two such plasmids (pT9E42, from transformation of WFD35, and pT11E60, from transformation of WFD36). Only one other Trp^- mutant, the indole-utilizing WFD135, could be complemented with the information cloned in these plasmids.

Similar cloning experiments with the more recently developed shuttle vector pWL102 (5) were used to analyze the remaining 19 indole- or anthranilate-utilizing Trp^- mutants. Wild-type DNA partially digested with *Msp* I or *Hin*PI was separately fractionated on sucrose gradients. DNA fragments 4–7 kb in size were purified and ligated into the unique *Cla* I site of pWL102. (The shuttle vector DNA was prepared from *H. volcanii* to protect ligated DNA from digestion when transformed into *H. volcanii* auxotrophs.) Recombinant plasmids isolated from prototrophic, mevinolin-resistant transformants shared some common restriction sites and could be used to define a 7-kb region in which these mutations lie (Fig. 1). In fact, three plasmids (pT1323-15, pT240-1, and pT186-4) that contain no extra sequences outside this 7 kb could transform 19 of the 20 indole- or anthranilate-utilizing mutants (all but WFD135; see above) to Trp^+, placing the gene for at least one component of the anthranilate synthase (EC 4.1.3.27) system and at least one other *trp* gene in this region.

Genetic Mapping of *trp* Genes. In preparing a bottom-up restriction map for *H. volcanii*, we have assembled a minimally overlapping set of 154 cosmids covering >95% of the genome (6). Cosmid DNAs (individually or in pools) can be used to transform *H. volcanii* auxotrophs to prototrophy (14). The cosmids cannot replicate in *H. volcanii*, and appear to rescue mutants by homologous recombination. Such cosmid transformation allows us to map mutations and to localize genes on the 3.8 million-base-pair (bp) genome to within 1%.

We used the 29 Trp^- auxotrophs to develop strategies for locating genes within the developing genome map. A pooling scheme (in which each of the 154 cosmids is represented in two separate pools) allowed us to deduce, from a single transformation experiment with 25 cosmid DNA pools and two control samples, the individual cosmid that bears the wild-type allele for any auxotrophic mutation. Of the five Trp^- mutants initially tested, two (WFD131 and WFD287) mapped unequivocally to cosmid 452. The remaining three (WFD261, WFD300, and WFD531) mapped to cosmids 488 and A159. These latter two cosmids overlap by 33 kb, and the mutations must lie within the overlap. We then transformed each of our 29 auxotrophs with cosmids 452 and A159. All mutants that grow only on tryptophan, as well as the indole-utilizing WFD135, could be transformed with cosmid 452, whereas the rest, which utilized indole or anthranilate, mapped to cosmid A159. These locations are separated by a minimum of 120 kb according to the physical mapping data, which has linked the cosmid set of 154 into 11 large map

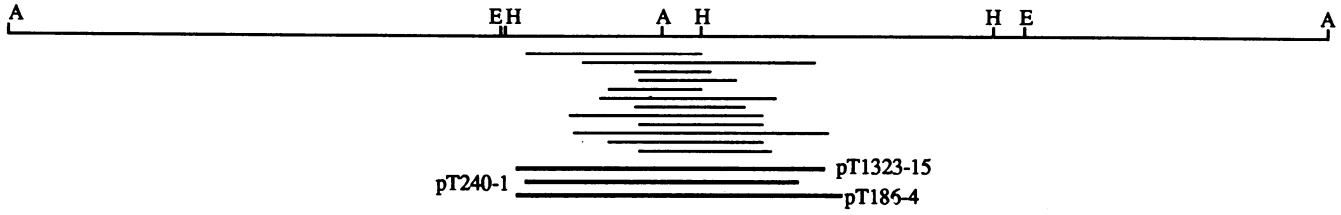

FIG. 1. Minimal region required for complementation determined (by aligning restriction maps of plasmid clones) for individual indole- and anthranilate-utilizing mutants. DNA inserts (*Msp* I or *Hin*PI partial digests) cloned in plasmids pT1323-15, pT240-1, and pT186-4 mentioned in the text and Table 1 are indicated. E, *Eco*RI; H, *Hin*dIII; A, *Apa* I. Distance between the *Eco*RI sites is 12.5 kb.

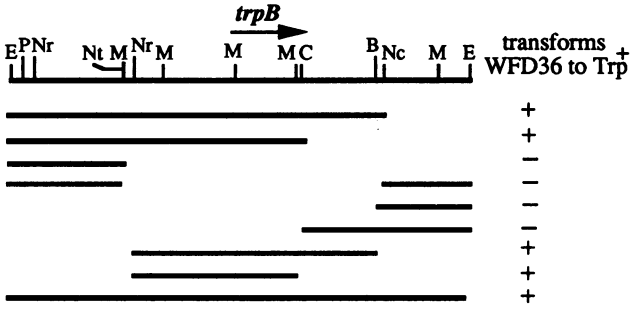

FIG. 2. Deletion analysis of pDWT1. Subclones and deleted forms of pDWT1, generated by standard methods, were tested for their ability to transform the Trp⁻ strain WFD36 to Trp⁺. E, *Eco*RI; P, *Pst* I; M, *Mlu* I; C, *Cla* I; B, *Bam*HI; Nc, *Nco* I; Nr, *Nru* I; Nt, *Not* I.

fragments (R. Charlebois, L. Schalkwyk, J. Hofman, and W.F.D., unpublished data). Similar mapping experiments have placed guanosine loci on cosmids near cosmids 488 and A159 and tyrosine loci on a cosmid close to cosmid 452.

Plasmid clones bearing *trp* genes can be labeled and used as probes against dot-blot filters bearing DNA from all 154 cosmids of the minimal set in an ordered array. This provides an independent method of assigning loci on the physical map and establishing linkage between markers. All cosmid assignments made by transformation could be confirmed by this procedure. Mutant WFD246, which reverted at too high a frequency to be mapped by transformation with cosmid DNA (since this is subject to restriction in *H. volcanii*), could only be mapped in this way. All Trp⁻ mutations so far obtained can be mapped to one of two unlinked regions of the *H. volcanii* genome using these independent physical and genetic methods.

Delimitation and Sequencing of *trpCBA* Genes. The ability to transform *H. volcanii* auxotrophs with small DNA fragments (14) allows more precise localization of genes within plasmid or cosmid clones. We made plasmid pT11E60 into a hybrid vector pDWT1 maintainable in *E. coli* by insertion of the pBR322 derivative pAT153 into its unique *Hin*dIII site. The 8-kb *Eco*RI fragment of pDWT1 (Fig. 2) was able to rescue all mutants absolutely requiring tryptophan and the indole-utilizing WFD135, thus placing the *trpB* gene and at least one other *trp* gene within this fragment (Table 1). Using subfragments of this 8-kb insert to transform WFD36 allowed us to identify a 2.8-kb *trpB*-containing region (Fig. 2). This region also transforms WFD135 to Trp⁺.

The nucleotide sequence of this and neighboring regions shows three open reading frames (ORFs) transcribed in the same direction (Fig. 3). These ORFs were named *trpC* (756 bp), *trpB* (1272 bp), and *trpA* (834 bp) based on their amino acid sequence similarities to eubacterial and yeast homologs (Fig. 4 and Table 2).

The deduced amino acid sequence of *H. volcanii* TrpB is about 45% identical to homologs in eubacteria, methanogen, and yeast (Table 2). The 423-residue *H. volcanii* TrpB sequence, although significantly larger than that of *E. coli* (397 amino acids), aligns coherently with eubacterial tryptophan synthase (EC 4.2.1.20) β-subunit sequences (16) and with the B-chain sequence derived from part of the yeast *trp5* gene (17) (Fig. 4). Extra residues at the C terminus account for some of the size difference between the eubacterial and the *H. volcanii* TrpB. Three insertions (of 2, 6, and 8 residues at positions 40, 263, and 367 of the *E. coli* sequence), introduced to optimize the sequence alignment, are found in or adjacent to regions predicted by the *Salmonella* three-dimensional structure (18) to be random coils. Half of these 16 inserted residues are acidic residues—not surprising because most proteins in halophilic archaebacteria are acidic (19). The nine extra residues found in *M. voltae* TrpB (at position 245 of *E. coli*/mh;6u) are not present in *H. volcanii*.

The catalytic residue Lys-87 (20) and the Gly-Gly-Gly-Ser-Asn stretch at positions 232–236 involved in coenzyme binding (7, 18) are conserved in *H. volcanii*, along with other residues involved in TrpB function (His-82, His-86, and Cys-230). The invariant tryptophan residue of unknown function at position 177 is also found (17).

The 834-bp ORF downstream from *trpB* encodes a 277-residue peptide resembling the *E. coli* TrpA sequence (Fig.

```
                                                                    -400 GGTCGGTCGGGACGGTCAGTCGATGGGGAC
CGACGAGAGGTCCGTCTGGCCCCGGTCGATGGCGGCCAACACGGCGTCGAGGTGGGCGTGGTCCGACTCGGAGTCCGCCGGCAGCGAGTCCGGAAACGAGACGTCGATGA
CCTTCCCGCTGACCACGCCGACCTGCACGGGGCGGCCGAGCGTGGACGATTCGCGCGCGTAGATTCCCGAGACGGGCATACGTTCGGCTCCCGTCGCCGGACGCTTGACC -151
              -130.                                  -100.                                  -70.                  -50.
CTTCCGGCGACGGCGTGCGGTTGGGGTGACTGCGCTCGGTTCGGGTGGCCGATACCTCGCCGTGCGCTCTCGCCCGCGCGCACGCGAGAGACGGTGAGTGTCCCCGCCGA -41
                                                                                                  trpC
CGTAAGCCTTATGTACAGATACGTGCGTTAGTGTACATCAATGAACGCTAGTGGAGACGAATTAGCTCCCGACGTGCGCGCCATCTTGGAGGCCGCGCGGGAGCGACCCG 70
GCGGGGAGACGCGCGTGTCGGTCGACGCGCGGTCGTTCCCCGAGGCCGTGGCCGAGACGGAGGCCGCGGGTCGGGTCCCCGTCATCGCCGAGGTGAAGCCCACGAGTCCG
ACGACCGAGGGCGTCCGCGAGGACGACCCGGTCGAACTGGCCCGCGAGATGGTCGCCGGCGGCGCGACGGCGCTGTCGGTCCTCACCGAACCCGAGCACTTCGGAGGCTC 290
CGCCGAGTCGCTGCGGCGCATCCGCGAGGCCGTCGACGTGCCCGTGCTCCGCAAGGACTTCATCATGAACGAGGCCCAACTGGATGTCGTCCAGTCCGACCTCGTGCTCC
TCATCGCGCGGTTCGTCGGCGAGGACCTCCCGGCCCTCGTCGAGGCGGCCCGCGACCGCGGCTTCCAGCCGCTCGTGGAGGTCCATACGCGCGAGGAACTCACGGCGGCG 510
CTCGCGGCCGGTGCCGACATCGTCGGCATCAACAACCGCGACCTCGGGAAGCTGGAAGTCGACCTCGGCACGTTCGAGGAACTCGCGCCCGAAGCGCCCGAGGACGTGCT
TCTCGTGGCCGAAAGCGGCGTGCAGACGGTCGACGACGCGCGGCGGATGCGCGAGGCCGGTGCCGACGCCCTCCTCGTCGGGACCGCCATCATGGACGGCGACGTGCGAC 730
AGAACACGGAGACACTCACACATGA                trpB
                          GCGCAGACGGCAAATTCGGCGACTACGGCGGACAGTACGTTCCCGAGGCACTCATGCCGGCCATCGAGGAACTGACCGACGCGT 840
ACGAGCGGTACGTCCTCGACAACGAAGACGGCTTCATGGACGACTTCCGCGCGCGACTGCGGGACTTCGGCGGGCGACCGACGCCCCTCCAGCGCGCGGACCGACTCTCG
GAGCGCTACGACCGCGAGGTCTACCTCAAGCGCGAGGACCTCCTCCACGGCGGCGCGCACAAGCTCAACAACGCGCTCGGGCAGGTCTTACTGGCGAAGTACATGGGCAA 1060
AGAGCGCATCATCGCCGAGACCGGCGCGGGCCAGCACGGCACCGCCACGGCGATGGCCTGCGCGCACCTCGATATGCCCTGCGAGATTTACATGGGCGAGCGCGACATCA
ACCGCCAGCGCCCCAACGTCTTCCGGATGAAGCTCAACGGCTCCGAGGTGAACCCCGTCACCGTCGGCCGCGGCACGCTCAAGGAGGCCATCTCCGAGACGATGCGCGAC 1280
TGGGAGACCAACGTCGAGGACACCCACTACGTCATCGGCTCCGTCGTCGGCCCGCACCCGTTCCCGAGCATGGTTCCGAGTCGGTCATCTCCGAGGAGGCCCG
CACGCAGGCCAGAGAGAAACTCGGCCGGCTCCCCGACGCCGTCGTCGCCTGCGCGGGCGGCGGCTCGAACACGATGGGCGCGTTCGCCGAGTTCGTCGACGACGAGGAGA 1500
CCGCGCTCTACGCCGTCGAGGCCGGCGGCTCGACGCTCGAAGTGGACGAGGAAGCCGGCGTCGCGCCCAACTCGGCGTCGCTCACGACCGGGTCGGAGGGCATTCTCCAC
GGCGCGCGCACCCGGCTCTTGCAGGACCGCGACGGCCAGATTATGGAGTCGCACTCGGTGTCGTCCGGCCTCGACTACGCCGGCGTCGGCCCCGAACTCGCACACCTCGT 1720
GGACACCGGCCGCGTCACCGCCGTCAACGTCGACGACGACGCGGCGCTGACCGCGTTCCACCGGCTCTCGCAGATGGAGGGCATCATCCCCGCCCTGGAGTCGGCCCACG
CGTTCGGCTACCTCGAAGTCTGGTCGTCTGGTCCCGACGCGCCCGACGCCGAGAACGCGGACGACCTCGGCGAGTACGTCGTGGTCAACGTTTCCGGTCGCGGCGACAAA 1940
GACCTCGAATCGGCCATCGAGGAGACCTACGAGCGCGACATCGATATCGCGCCGAACATGGACGAGTTCACGGGGGGCCTGTGA                trpA
                                                                                                  TGTCGCTCGAAGACGCCTTCTCCGAC 2050
GGCCCGGCGTTCGTCCCCTATCTTGCCGCCGGCGACCCCGACTACGAGTCCTCGCTCGAATACGTCGAGGCGCTCGAACGCGGCGGCGCGGACGTCATCGAACTCGGACT
TCCGTTCTCGGAGCCCATCGCCGAGGGGCCGACCATCCAGAACGCGGTTGTTCGGTCGCTCGAAGGCGGCATGACGCCGACGCGCTTTTTCGAGTTCGTCGAGGACCTCG 2270
ACGTGTCGGTGCCGCTGGTCTGTATGAGGTACTACAACCTCATCTATCGCTACGGCGATGAACCCGGACCGCGGCCGTTCGTCGAGAAGGCGGCGGAAGTCGGTATCGAG
GGCTTCGTCGTCCCCGACCTGCCGGCCGAGGAGGCCGGCCCGCTCCGCGAGGCCTGCGACGAGTTCGGCCTCGACCTCGTGTTCATCGTCGCGCCGACGACCCGCGGCGA 2490
GCGCCTCGACCGAATCATGGAACAGGTCTCGGGCTACGTCTACGTGCAGGCGCGCCTCGGCACGACGGGCGCGCAGTCGAGCGTCTCCGACCAGACCGACTCGTCGCTCG
AACGACTCACCGACTACGACGTGCCCAAGGCGGTCGGCTTCGGAATCAGCGACGGCGACCACGCCGAGCGCATCGTCGCCAGCGGTGCCGACGGCATCATCGTCGGCAGC 2710
GCGCTCGTCGACATCGTCGCCGAGGGCCACGAGAACGGCGACGACGCCGAAACGGTCGCGGACCGACTCGAAACGCTCGCCCGCGAACTCGAAGACGGTGCGGTAGCGGG
CGCGTCGCAACGCCCACCGCATCCGGAACGCACATAA
                                                    CTGCTCTGCCACCCTTTCTCATACATGAACACAGACGTCGGACTCTCCGCACGACTCGACCGCATTTCCACAG 2930
ACGGGCGATACCTCATCGTCCCGATGGACCACGGCATCACCCTCGGCCCGGTCACGGGCCTCGTCGATATCGAATCCACTATCGACGGCATCACGCGCGGCGGCCGTGAC
GCCGTCCTCACGCACAAGGGGACCGCGCCGCGCGTCCACCCGAACAAAAACGGCAAAGGCTACATCGTCCACGTAAACGGCTCGACGGACATCGGCCCGGACGAAAACGA 3150
```

FIG. 3. DNA sequence of the *trpCBA* region. Nucleotides comprising overlaps between reading frames are shown as part of the upstream ORF and are boxed. Arrows indicate repeated sequences.

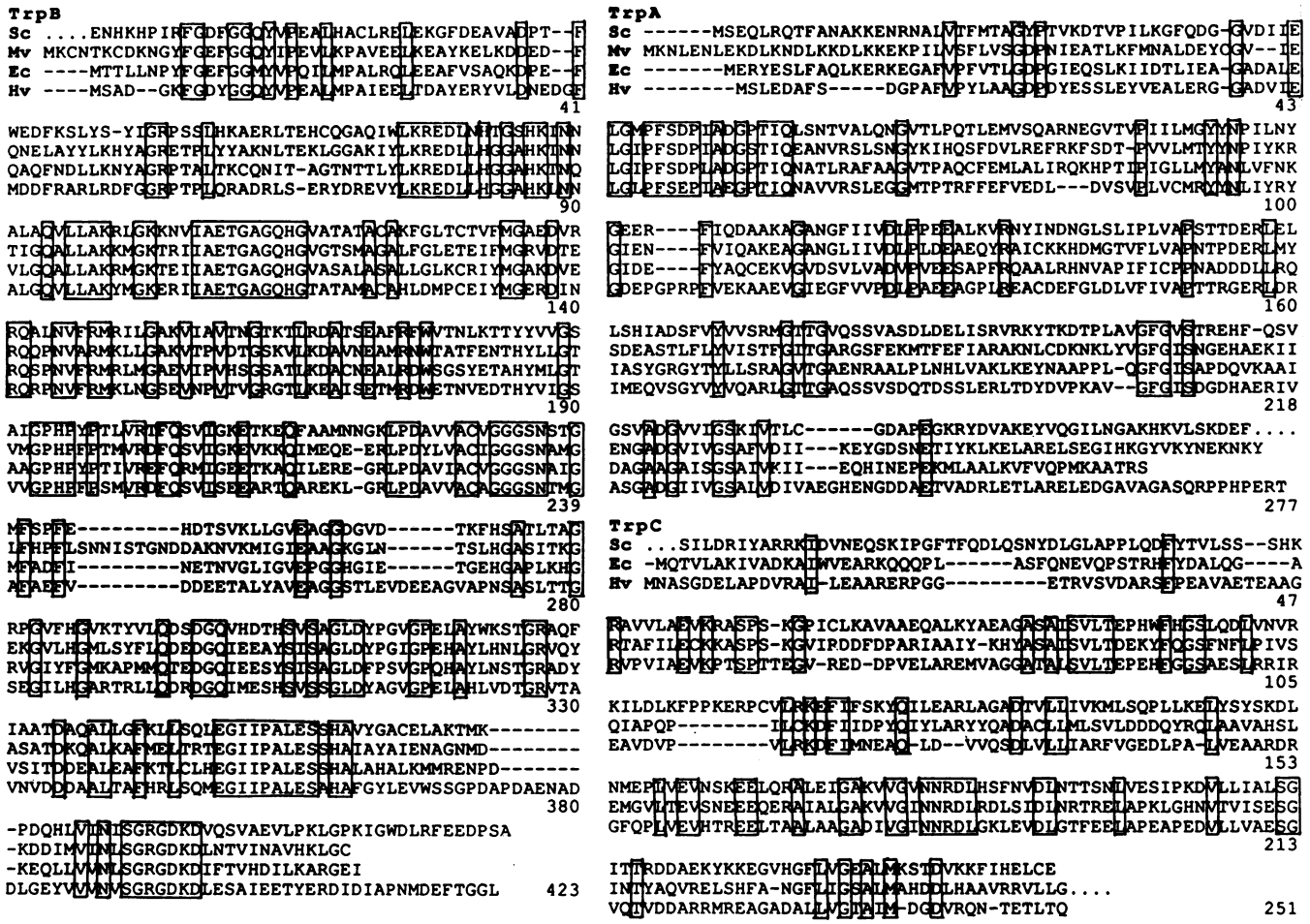

FIG. 4. Alignment of amino acid sequences for TrpB, TrpA, and TrpC from yeast (Sc), *M. voltae* (Mv), *E. coli* (Ec), and *H. volcanii* (Hv). Sequences were aligned by the MULTALIN method of Corpet (11). Very few adjustments were required to conform the alignments to those reported by other groups (8, 15). Conserved residues are boxed. The number at the end of each block refers to the position of the *H. volcanii* sequence.

4). As in *E. coli*, the start codon (AUG) of *trpA* overlaps the stop codon (UGA) of *trpB* by one nucleotide (UGAUG). In contrast, the *trpB* and *trpA* genes in the methanogen are separated by a 37-bp A+T-rich region (8).

Overall sequence conservation for TrpA is significantly lower than for TrpB (Table 2). The *H. volcanii* TrpA is about 30–40% identical to all TrpA sequences we have compared. Important residues (Phe-22, Glu-49, Try-175, Thr-183, Gly-211, Gly-234, and Ser-235) defined by missense mutations and second-site reversion in *E. coli* (7, 20, 21) are all found in the *H. volcanii* sequence except that Tyr-16 substitutes for Phe-22. Leu-177 is replaced by Asn-172 in *H. volcanii*; replacements by isoleucine and valine at this position have been reported previously in other organisms (8, 17). The

Table 2. Conservation of TrpC, TrpB, and TrpA sequences between *H. volcanii* and other organisms

Organisms compared	Identical amino acids, %		
	TrpC	TrpB	TrpA
E. coli	32	45	32
S. cerevisiae	36	45	33
P. aeruginosa	38	49	37
M. voltae	—	47	40
B. subtilis	32	45	30

Amino acid sequences used in multiple alignments as in Fig. 4 were further compared pairwise to *H. volcanii* sequences for identical residues. *P. aeruginosa*, *Pseudomonas aeruginosa*; *B. subtilis*, *Bacillus subtilis*; *S. cerevisiae*, *Saccharomyces cerevisiae*.

Gly-Phe-Gly-Ile stretch that contacts the substrate in *E. coli* (7) is located, by sequence alignment, at positions 205–208 of the *H. volcanii* TrpA.

In the α subunit of the *Salmonella* tryptophan synthase, Glu-49 and Asp-60 serve as proton donors/acceptors in the cleavage of indole-3-glycerol phosphate. Replacement of Asp-60 by asparagine, alanine, or tyrosine destroys α-chain activity, but replacement by glutamic acid permits partial activity (20). In fact, glutamic acid (Glu-54) is the residue found at the corresponding position in *H. volcanii*.

Immediately upstream of *trpB* is a 756-bp ORF that overlaps *trpB* by four nucleotides (AUGA, Fig. 3). The deduced polypeptide (251 amino acids) has stretches of conserved sequences common to the indole-3-glycerol-phosphate synthase (EC 4.1.1.48) domain of yeast (22) and to eubacterial (15, 16) TrpC sequences (Fig. 4 and Table 2). Residues that may contribute to catalytic activity of *E. coli* TrpC are mostly conserved in *H. volcanii*.

Codon usage of the *trpC*, *trpB*, and *trpA* genes is very much biased to codons with cytidine or guanosine in the third position, reflecting the high G+C content of these genes (68–71 mol %). This is consistent with the >65 mol % G+C content estimated for halobacterial genomes (23).

Transcription and Translation Signals. The fact that *trpC*, *trpB*, and *trpA* overlap by their stop and start codons suggests that these genes are transcribed as a unit from the same promoter. Approximately 30 bp upstream from the start codon (AUG) of *trpC* is the putative halobacterial promoter sequence TTATGTA (23), followed by two nearly perfect

direct repeats 10 bp apart (see Fig. 3). The sequences GTGAT (at the *trpB/trpA* junction) and GGAG (near the end of *trpC*, 11 bp upstream of *trpB*) are both complementary to the 3′ terminus of the small subunit ribosomal RNA of *H. volcanii* (3′-UCCUCCACUA . . .) and may serve as ribosome-binding sites.

The sequence within 100 bp further upstream from the putative promoter can potentially be folded into stable secondary structures. Remarkably, the sequence between −76 and −95 can form alternate stable stem-and-loop structures with the adjacent sequences immediately upstream or downstream (from position −127 to −112 and from −67 to −53). Site-specific mutagenesis experiments are required to assess any role of these sequences in gene expression or its control.

Evolutionary Considerations. The *H. volcanii* TrpB and TrpA amino acid sequences are as different from all sequences we have compared them to (including that of the methanogen) as the eubacterial sequences are from the eukaryotic ones (Table 2). The same observation applies to TrpC, although *Methanococcus* sequences are not available for comparison. The rate of amino acid substitution outside the invariable positions is too high to allow the use of these sequences for assessing phylogenetic relatedness of organisms. However, the existence of conserved regions in TrpC, TrpB, and TrpA indicates the presence of these polypeptides before the divergence of archaebacteria, eubacteria, and eukaryotes.

As in eubacteria, the two domains of *H. volcanii* tryptophan synthase are encoded by transcriptionally linked but separate genes. The *trpB* and *trpA* genes are preceded by *trpC*—an arrangement unique to *H. volcanii*. Methanococcal and eubacterial *trpB* and *trpA* genes either follow *trpF* or exist as a unit alone (8, 15). The *trpC* gene is either fused to or transcriptionally linked to *trpF*, *trpG*, or *trpD* (7, 15, 24).

The difference in gene arrangement between *H. volcanii* (*trpCBA*) and *M. voltae* (*trpFBA*) is perhaps not surprising, since no unified arrangements are common to all eukaryotes or all eubacteria (except for the linkage of *trpB* and *trpA* in prokaryotes). Diversity is apparent even within the γ purple bacteria: *E. coli* has a single *trp* operon, but the *trp* genes of *P. aeruginosa* exist in four transcriptional units and respond to three distinct regulatory mechanisms (15).

Nevertheless, *trp* genes tend to be fused or transcriptionally linked. Unless linkage provides tremendous selective advantage, it is difficult to envision seven separate genes repeatedly seeking out one another to form fusion units and operons; thus, the ancestral set of *trp* genes must have been linked and since then has been shuffled, fused, and rearranged in ways only restricted by the needs of the organism. It is important to find out how generally archaebacterial biosynthetic pathway genes show such operon-like clustering. Genetic approaches such as those developed here should make this possible.

We thank R. L. Charlebois for providing the cosmid set, R. K. Conover for expert help with computer analyses; S. W. Cline for frozen spheroplasts; and A. Alberts for mevinolin. W.F.D. is a Fellow of the Canadian Institute for Advanced Research. This work was supported by grants from the U.S. Office of Naval Research (N00014-88-J-1030) and the Canadian Medical Research Council (MT4467).

1. Woese, C. R. & Fox, G E. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5088–5090.
2. Woese, C. R. (1987) *Microbiol. Rev.* **51,** 221–271.
3. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579.
4. Charlebois, R. L., Lam, W. L., Cline, S. W. & Doolittle, W. F. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 8530–8534.
5. Lam, W. L. & Doolittle, W. F. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5478–5482.
6. Charlebois, R. L., Hofman, J. D., Schalkwyk, L. C., Lam, W. L. & Doolittle, W. F. (1989) *Can. J. Microbiol.* **35,** 21–29.
7. Crawford, I. P. (1989) *Annu. Rev. Microbiol.* **43,** 567–600.
8. Sibold, L. & Henriquet, M. (1988) *Mol. Gen. Genet.* **214,** 439–450.
9. Henikoff, S. (1984) *Gene* **28,** 351–359.
10. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
11. Corpet, F. (1988) *Nucleic Acids Res.* **16,** 10881–10890.
12. Queen, C. & Korn, L. (1984) *Nucleic Acids Res.* **12,** 581–599.
13. Mevarech, M. & Werczberger, R. (1985) *J. Bacteriol.* **162,** 461–462.
14. Conover, R. K. & Doolittle, W. F. (1990) *J. Bacteriol.* **172,** 3244–3249.
15. Essar, D. W., Eberly, L. & Crawford, I. P. (1990) *J. Bacteriol.* **172,** 867–883.
16. Yanofsky, C., Platt, T., Crawford, I. P., Nichols, B. P., Christie, G. E., Horowitz, H., van Cleemput, M. & Wu, A. M. (1981) *Nucleic Acids Res.* **9,** 6647–6668.
17. Zalkin, H. & Yanofsky, C. (1982) *J. Biol. Chem.* **257,** 1491–1500.
18. Hyde, C. C., Ahmed, S. A., Padlan, E. A., Miles, E. W. & Davies, D. R. (1988) *J. Biol. Chem.* **263,** 17857–17871.
19. Eisenberg, H. & Wachtel, E. J. (1987) *Annu. Rev. Biophys. Biochem.* **16,** 69–92.
20. Hyde, C. C. & Miles, E. W. (1990) *Biotechnology* **8,** 27–32.
21. Yanofsky, C. & Crawford, I. P. (1972) in *The Enzymes*, ed. Boyer, P. D. (Academic, New York), 3rd Ed, pp. 1–31.
22. Zalkin, H., Paluh, J. L., van Cleemput, M., Moye, W. S. & Yanofsky, C. (1984) *J. Biol. Chem.* **259,** 3985–3992.
23. Brown, J. W., Daniels, C. J. & Reeve, J. N. (1989) *CRC Crit. Rev. Microbiol.* **16,** 287–334.
24. Yanofsky, C. (1984) *Mol. Biol. Evol.* **1,** 143–161.