The tryptophan repressor sequence is highly conserved among the *Enterobacteriaceae*

Dennis N. Arvidson*, Cindy Grove Arvidson¹, Catherine L. Lawson², Jami Miner, Curtis Adams⁺ and Philip Youderian

California Institute of Biological Research, 11099 N. Torrey Pines Road, La Jolla, CA 92037,

¹Department of Microbiology and Immunology, Oregon Health Sciences University, Portland,

OR 97201 and ²Biology Department, Brookhaven National Laboratory, Upton, NY 11973, USA

Received February 11, 1994; Revised and Accepted March 28, 1994

GenBank accession nos L26582, L26583

ABSTRACT

Tryptophan biosynthesis in *Escherichia coli* is regulated by the product of the *trpR* gene, the tryptophan (Trp) repressor. Trp aporepressor binds the corepressor, L-tryptophan, to form a holorepressor complex, which binds *trp* operator DNA tightly, and inhibits transcription of the tryptophan biosynthetic operon. The conservation of *trp* operator sequences among enteric Gram-negative bacteria suggests that *trpR* genes from other bacterial species can be cloned by complementation in *E. coli*. To clone *trpR* homologues, a deletion of the *E. coli trpR* gene, $\Delta trpR504$, was made on a plasmid by site-directed mutagenesis, then crossed onto the *E. coli* genome. Plasmid clones of the *trpR* genes of *Enterobacter aerogenes* and *Enterobacter cloacae* were isolated by complementation of the $\Delta trpR504$ allele, scored as the ability to repress β -galactosidase synthesis from a prophage-borne *trpE-lacZ* gene fusion. The predicted amino acid sequences of four enteric TrpR proteins show differences, clustered on the backside of the folded repressor, opposite the DNA-binding helix-turn-helix substructures. These differences are predicted to have little effect on the interactions of the aporepressor with tryptophan, holorepressor with operator DNA, or tandemly bound holorepressor dimers with one another. Although there is some variation observed at the dimer interface, interactions predicted to stabilize the interface are conserved. The phylogenetic relationships revealed by the TrpR amino acid sequence alignment agree with the results of others.

INTRODUCTION

The tryptophan (Trp) regulon is one of the most intensively studied procaryotic model systems for global gene regulation (1–3), repressor protein structure and function (4, 5), and molecular evolution (6, 7). Trp holorepressor recognizes at least

five operators on the *E. coli* genome to repress the transcription of the *trpEDCBA* (8), *aroH* (9, 10), *mtr* (11–13), *aroL* (14), and *trpR* (15) operons. The *E. coli* operators bound by Trp repressor are conserved in sequence but not with respect to their positions relative to transcription initiation sites, and are an example of convergent evolution (6).

The study of interspecific hybrids shows that the Trp repressor–operator interaction is conserved among the Gram-negative enteric bacteria. The *Shigella dysenteriae trp* operon is regulated by *E. coli* Trp repressor (16, 17), and the *E. coli trp* operon is regulated normally by *Shigella dysenteriae* TrpR (17). An F'-borne *E. coli trpR* gene regulates *trp* operon transcription in *trpR*[–] *Salmonella typhimurium* and *Enterobacter aerogenes* hosts (17). Expression of the *E. coli trp* operon from an F' element in *S. typhimurium* (18), *E. aerogenes*, *Serratia marcescens*, and *Proteus mirabilis* is regulated in response to tryptophan, as in *E. coli* (17).

DNA sequences have been determined for the *trp* operators of *E. coli* (19), *S. dysenteriae* (20), *S. typhimurium* (21), *E. aerogenes* (22), *Citrobacter freundii* (23, 24), and *S. marcescens* (25). The central 18 base pairs of these operators are identical except for the *S. dysenteriae* and *S. marcescens* operators, which differ by single base pair changes. Similarities among *trp* operator sequences extend outward from this central region into flanking binding sites for repressor (26). Because these flanking sites contribute significantly to repression in an *E. coli* host (27, 28), it is likely that Trp repressors from different enteric species will share not only the ability to bind tryptophan and similar operators, but also the ability to bind in tandem to sites within these compound operators.

Skrypka and Somerville have cloned and sequenced the *S. typhimurium trpR* gene (29). The predicted amino acid sequences of the *S. typhimurium* and *E. coli* encoded Trp repressors differ for only three of 107 residues. We constructed an *E. coli* strain with a small, defined deletion of *trpR* to clone *trpR* genes from the more distantly related bacteria, *E. aerogenes* and *E. cloacae*, by complementation. Comparison of the four

*To whom correspondence should be addressed

⁺Present address: Department of Microbiology, University of Washington, Seattle, WA 98195, USA

known enteric *trpR* genes shows that the Trp repressor coding sequences are highly conserved. The conserved sequences reveal common structural features of the *trpR* genes, suggest the possible occurrence of insertion-deletion events in their evolution, and allow us to model their evolution from a common, ancestral *trpR* gene.

MATERIALS AND METHODS

Bacteria, plasmids and phage

E. aerogenes (ATCC 13048) and *E. cloacae* (ATCC 23355) were from the University of Southern California Instructional Microbial Culture Collection. *E. coli* strains are derivatives of K-12 and are listed in Table 1. Plasmids included the ColE1-derived vectors pZ152 (30), pKRS101, which carries the *trp* promoter/operator region (31), pZ178, which carries a promoter-less *lacZ* gene and the M13 origin of replication (Berman, M.L., and Zagursky, R., unpublished), pBluescript SKII⁺ and KSII⁺ DNA (Stratagene), and pRPG5 (15), which carries the *E. coli trpR* gene and flanking sequences. The region of *E. coli* chromosomal DNA with the deletion of *trpR* was subcloned into plasmid pACYC184 (32). The specialized transducing phage λ CLG145 (Fig. 2), carrying a *trpE-lacZ* fusion, was constructed from λ RS45 (33). Single-stranded plasmid templates used in sequencing reactions were prepared after infection of *E. coli* hosts with M13 VCS helper phage (Stratagene).

Media and general methods

LB (rich) and M9 (minimal) media were used for routine growth of cells (34); media, agar, and MacConkey agar were from Difco. Antibiotics and nutritional supplements were from Sigma. X-gal (5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside) was from Stratagene. General cloning techniques were as described (35). Restriction enzymes (Stratagene or New England Biolabs) were used under the conditions specified by the supplier. DNA-DNA hybridization (Southern blot) experiments were performed as described (36).

Construction of *E. coli* strains with a defined *trpR* deletion

The plasmid with $\Delta trpR504$ (pDNA18, Fig. 1) was constructed in several steps. First, pDNA11 was constructed by subcloning the 1283 bp *Bam*HI fragment carrying *trpR* from pRPG5 (15) into pZ152 (30). Second, a synthetic oligonucleotide 5' GCGCCACGGAATGGGGA/TCCTACAAATACCCGTA, that corresponds to the sense strand of *trpR* and fuses base 283 to base 788 (GSDB, J01715 V00369) generating a *Bam*HI site at the deletion joint, was used to prime the synthesis of the second strand from single-stranded pDNA11 (37). The mutagenized plasmid DNA was used to transform *E. coli* X90/*F'**lacI*^{Q1} and the sizes of plasmids carried by ampicillin-resistant transformants were screened. Plasmids that acquired the deletion were identified by cleavage with *Bam*HI, and the partial 779 bp *Bam*HI fragment from one plasmid with the desired deletion junction, pDNA14, was subcloned into the *Bam*HI site of pACYC184 (32) to construct pDNA18.

To screen for the loss of repressor activity in a host transformed with pDNA18, we constructed a specialized transducing derivative of phage λ , λ CLG145, which carries the *trp* promoter/operator region and an in-frame fusion of the 31st codon from *trpE* with the 8th codon of *lacZ* (Fig. 2). The 490 bp *Sau*3A1 fragment with the *trp* promoter/operator region, *trpL*, and start of *trpE* from plasmid pKRS101 (31) was cloned into the unique

*Bam*HI site of plasmid pZ178 (Berman, M.L., and Zagursky, R., unpublished), yielding pCLG100. The gene fusion was crossed from plasmid pCLG100 to the specialized transducing phage, λ RS45 (33) to yield λ CLG145, identified as a blue plaque-forming phage on a lawn of *E. coli* P90C cells grown with 40 μ g/ml X-gal.

To cross the plasmid-borne $\Delta trpR504$ mutation onto the chromosome, pDNA18 was transformed into the *trpR*⁺ Δlac host, P90C(λ CLG145), and transformants were plated on MacConkey lactose plates with chloramphenicol (30 μ g/ml). After one day of incubation at 37°C, the transformants formed white colonies; after three days, the colonies developed multiple red papillae. In contrast, colonies formed by transformants with pRPG18, an otherwise isogenic plasmid with the *trpR* gene, or

Table 1. Bacterial strains

Strain	Genotype	Reference
X90	F' <i>lacI</i> ^{Q1} (<i>recA argE-am thi</i> Δ (<i>proAB lac</i>), Nal ^R , Rif ^R)	(85)
P90C	F ⁻ Δ (<i>lac-pro</i>) <i>ara thiI</i>	(86)
SP516	HfrH Δ 37-1(<i>deoD-trpR</i>) Δ (<i>gal-bio</i>)76 <i>relA1 spoT1</i> <i>thiI deoC76</i> λ^- PO1 <i>zjj::Tn10</i>	(39)
RH339	<i>recA56 srl::Tn10 uncH239 bglR thiI relI</i>	(43)
CG107	F ⁻ Δ (<i>lac-pro</i>) <i>ara thiI</i> $\Delta trpR504$ (λ CLG145)	this paper
CG108	F ⁻ Δ (<i>lac-pro</i>) <i>ara thiI</i> $\Delta trpR504$ <i>zjj::Tn10</i> (λ CLG145)	"
DA450	F ⁻ Δ (<i>lac-pro</i>) <i>ara thiI</i> $\Delta trpR504$ <i>zjj::Tn10</i>	"
CG103	F ⁻ Δ (<i>lac-pro</i>) <i>ara thiI</i> $\Delta trpR504$ <i>recA56</i> <i>srl::Tn10</i> (λ CLG145)	"
CG101	F ⁻ Δ (<i>lac-pro</i>) <i>ara thiI</i> Δ 37-1(<i>deoD-trpR</i>) <i>zjj::Tn10</i>	"

Figure 1. Plasmid constructions. The *Bam*HI fragment from pRPG5 was subcloned into plasmid pZ152, and the *trpR* gene was deleted from this recombinant, pDNA11, by site-directed mutagenesis. A partial *Bam*HI fragment containing the $\Delta trpR504$ allele was subcloned into pACYC184, to make pDNA18.

CTATTTATCGTTAAGC, and 5'
CGCAGCAGCGTCTACCATC (*E.aerogenes*); 5'
GAGTACAAGAGGCAGC, 5' CGCTGGCTCATTTCGC, and
5' CCGTAAAGAAGGGGCAATTG (*E.cloacae*). Primary
nucleotide sequence data have been deposited in the Genome
Sequence DataBase of the Los Alamos National Laboratory
(GSDB accession numbers L26582, *E.aerogenes* and L26583,
E.cloacae). The GSDB accession number for *S.typhimurium trpR*
is L13768 (29).

RESULTS AND DISCUSSION

Construction of a defined *trpR* deletion in *E.coli*

Escherichia coli strain CG103 carries a small 504 base pair deletion of the *trpR* gene, $\Delta trpR504$, which extends through the *trpR* promoter and terminator, from -45 to +459 with respect to the identified startpoint of *trpR* transcription (15). To construct $\Delta trpR504$, the deletion junction was synthesized as an oligonucleotide primer for the site-directed mutagenesis of a *trpR*⁺ plasmid (37). A plasmid with $\Delta trpR504$ was introduced into a *trpR*⁺ host, and bacteria that acquired $\Delta trpR504$ as a chromosomal marker were screened as recombinants that express a single copy *trpE-lacZ* reporter operon constitutively (see Materials and Methods).

The growth rate of *E.coli* is reduced by deleting genes near *trpR*. Otherwise isogenic strains with the wild-type or $\Delta trpR504$ alleles grow faster and achieve a higher final cell density in supplemented minimal media (87) than a strain with the $\Delta 37-1$ allele (a larger deletion of the *trpR* gene, which extends into the neighboring *deoD* gene). The three strains have identical growth rates in rich (LB) medium (data not shown).

Cloning and sequencing *Enterobacter trpR* genes

E.coli strain CG103 carries $\Delta trpR504$ and an integrated λ prophage with a *trpE-lacZ* gene fusion which expresses β -galactosidase constitutively. Plasmid libraries with *Enterobacter aerogenes* and *Enterobacter cloacae* inserts were transformed into CG103 and *trpR*⁺ transformants were identified as white colonies (Lac⁻ phenotype) on MacConkey lactose indicator plates. Complementing inserts were subcloned into plasmids pBluescript KSII⁺ and SKII⁺, and sequenced. Comparison of the DNA sequences of the clones with the sequences of the *E.coli* (15, 47) and *S.typhimurium* (29) *trpR* genes established the location of *trpR* within each fragment and permitted their alignment by inspection.

Tryptophan repressor sequences

Figure 3 shows that amino acid sequences for Trp repressor are highly conserved among *E.coli*, *S.typhimurium*, *E.aerogenes*, and *E.cloacae*. Amino acid differences are observed at only 21/107 residues and 17/21 cluster in three regions of the protein: four residues near the amino-terminus (2, 4, 9, and 11), seven of the nine residues 24–32, located at the carboxyl-terminus of α -helix A and the start of the turn between α -helices A and B (24, 26, 27, 28, 29, 31, and 32); and six of the nine residues 100–108, located at the carboxyl-terminus of α -helix F and the flexible carboxyl-terminus of the protein (all but 101, 104 and 105).

Trp repressor is a small, symmetric homodimeric protein. X-ray and NMR analyses of the *E.coli* TrpR dimer show that each subunit is folded into six α -helices (A–F), joined by short surface turns. Four α -helices (A, B, C, and F) comprise the central dimerization core. These helices adopt the same conformation

in the different crystal forms (48–50) and determine the least mobile portion of the protein in solution (51). α -helices D and E form the helix-turn-helix DNA-binding substructure of the TrpR protein. In solution, α -helices D and E of aporepressor are poorly defined (52), but become increasingly more ordered with the binding of tryptophan, then operator DNA, evidenced by an increase in both the number and intensities of observable NOEs (nuclear Overhauser effects) between proton pairs (5, 53). In the crystal structure of aporepressor (49), the DNA-binding helices are positioned about 25 Å apart. In both trigonal (48) and orthorhombic (50) crystals of holorepressor, the DNA-binding helices are further apart (34 and 30 Å, respectively), presumably allowing the repressor to contact successive major grooves of *trp* operator DNA (54). Comparison of the four enteric TrpR sequences reveals only one difference in the DNA-binding helices, the conservative Ala→Ser change at position 91. The side chain of this residue projects from the surface of α -helix E tangent to the DNA-binding interface, and therefore is not likely to be involved in DNA binding.

The planar indole ring of the bound L-tryptophan in holorepressor is sandwiched between Arg54 (α C) and Arg84 (α E). Ile57 and Val58 (α C) line the bottom of the corepressor binding pocket. The solvent accessible edges of the pocket are defined by contacts between the α -amino group of bound L-tryptophan and main chain carbonyls of Asn30', Leu41', and Leu43' (from the opposing subunit, α B'), and between the L-tryptophan carboxylate and Arg84 (α E) (49, 53, 55). The N ϵ of the tryptophan's indole is also exposed to solvent. Glycine at position 85 (α E) permits the helix-turn-helix structure to accommodate the corepressor ligand; in other helix-turn-helix

Figure 4. Schematic diagram of the Trp repressor dimer showing positions of non-conserved residues on one subunit (white coil). The view is of the 'back side' of the repressor dimer, i.e., the face opposite that which interacts with DNA. The model used for the drawing is one *trp* repressor dimer of the tandem complex crystal structure (PDB code 1TRR, 61). Each non-conserved residue is represented by a black stick model of the side-chain and is labelled with the residue number. In a few cases (2, 29, 91), the side-chain is not visible; these residues are represented by residue numbers only. The carboxyl-terminal residues 105 to 108 are not present in 1TRR, so are represented by the label 'C-term'. The other subunit (gray coil) and the two L-tryptophan corepressor ligands (gray stick model) are shown for reference. Labels and coils were drawn with exaggerated perspective: smaller numbers and thinner coils are in back, larger numbers and thicker coils are in front. The diagram was generated with MOLSCRIPT (88).

proteins, this residue is often tryptophan (56). Only one amino acid sequence difference among the four enteric repressors affects a residue implicated in corepressor binding. Residue 58 is isoleucine in the two *Enterobacter* repressors; *E.coli* and *S.typhimurium* repressors have valine at this position. In *E.coli* Trp repressor, the side chain of Val58 is proximal to the 5-position of the bound L-tryptophan indole ring (49, 53, 57, 58). A mutant *E.coli* repressor with this conservative change is comparable in activity to wild-type *E.coli* TrpR *in vivo*, and has an altered specificity of corepressor recognition (57). Ile58 repressor is activated more poorly than the wild-type by the toxic analog, 5-methyltryptophan (5-MT). 5-MT can substitute for L-tryptophan as corepressor *in vivo* (38), and is bound by wild-type *E.coli* TrpR with a higher affinity *in vitro* (58, 59). This analog is toxic to both *E.coli* and *S.typhimurium*, either because incorporation of 5-MT into essential proteins renders them inactive, or because 5-MT both causes feedback inhibition of anthranilate synthase and acts as corepressor, the combined effect of which is tryptophan starvation (60). The substitution of Ile at position 58 of the *Enterobacter* Trp repressors may render them less sensitive to activation by naturally occurring 5-substituted indole analogs and confer a selective advantage in particular ecological niches.

The majority of amino acid differences among the four proteins affect residues with solvent exposed side chains located on the surface of the protein opposite the DNA-binding interface (Fig. 4). In the crystal structures of aporepressor and holorepressor, 14 residues of the protein are extensively involved in intersubunit interactions (residues 18, 19, 21, 23, 25, 30, 34, 41, 42, 51, 54, 55, 96, and 99). A surface area greater than 55 Å² of each of these residues is buried in the second subunit of the interlocked *E. coli* TrpR dimer (49). Each of these residues is conserved. Residues 26, 27, 28, 29, 32, 56, 100, 103, and 106 are also involved in the dimerization interface and are not conserved. However, the maximum surface area buried in the second subunit is 40 Å² for this group (49).

Differences in the carboxyl-terminus of α -helix A are not found at residues Leu25 and Tyr30, with side chains that contribute extensively to the solvent-excluded dimer interface. Other residues in this region are less deeply buried in the opposite subunit, but also contribute to dimerization. Asn28 and Ala29 form part of a pocket for Trp99' (from the opposite subunit) in

the *E. coli* TrpR structures. Apparently, the exact structure of this pocket is not critical, as conservative amino acid differences are found at these positions. Residues Asp24 and Lys27 form an intrasubunit salt bridge in *E. coli* TrpR protein. The aliphatic portion of Lys27 lies over the Tyr30' phenyl ring. In the protein sequence alignment (Fig. 3) residues 24 and 27 change as a pair. In *E. cloacae* the residues are Glu24 and Arg27, longer side chains with the same potential charges. In *E. aerogenes* the residues are Ala24 and Gln27, a non-polar and an uncharged side chain, respectively; thus, a salt bridge may no longer be required. In each case the interaction between the aliphatic portion of residue 27 and the Tyr30' phenyl ring is preserved. Residues 31 and 32, which are solvent-accessible in the repressor dimer, show nonconservative amino acid changes.

Residues 7, 61, 62, 68, 86, 87, 89, 90, 97 and more indirectly 2, 45, and 46 are thought to be involved in interactions between two tandemly bound dimers of *E. coli* Trp repressor (61). Each of these residues is conserved, except residue 2, which is Ala in *E. coli* and *S. typhimurium*, but is Thr in *E. aerogenes* and *E. cloacae*. The amino-terminal methionine of the *E. coli* TrpR polypeptide is not present in the mature folded protein (15). The free amino-terminal amino group of Ala2 of *E. coli* TrpR is involved in stabilizing a conformation of Asp46' that avoids a steric clash with Lys90" (from the neighboring tandem TrpR dimer, 61). As Met is probably efficiently processed from proteins having Thr as the second amino acid residue in the Enterobacter (as it is in *E. coli*, 62), this amino acid difference (Ala2 vs Thr2) is not likely to affect tandem binding.

The structure of the carboxyl-terminus of *E. coli* TrpR protein is poorly defined in both X-ray and NMR structures of the protein (5, 54). The variation in sequence of the carboxyl-terminus is due, in part, to insertion-deletion events in the end of the *trpR* gene (Fig. 5). Residues 106–108 are the least conserved contiguous residues of the protein. In a separate study we have shown that a mutation that changes Glu106 to Lys has little effect on *E. coli* TrpR activity *in vivo* (unpublished). Nonconservative changes are also found at residue 102, whereas conservative changes occur at position 103. The conservation of Leu104 and Leu105 within this highly variable region suggests that these residues are important for TrpR folding or function. Leu104 is partially buried in the opposite subunit of the TrpR dimer. Leu104 and Leu105 are involved in an intrasubunit interaction that helps

Figure 5. Amino acid and nucleotide sequences of the *E. coli*, *S. typhimurium*, *E. aerogenes*, and *E. cloacae* *trpR* 3' flanking region. The putative amino acid sequence of a deduced ancestral TrpR is shown above its nucleic acid sequence and the residues are numbered. The putative amino acid sequence of a deduced ancestral URF is shown below its nucleic acid sequence. The arrows show the point of insertion for the underlined sequences in the *Escherichia coli* and *Enterobacter cloacae* genes. TrpR codons are separated by vertical lines. URF codons are separated by colons. TrpR stop codons are shown in bold typeface. URF stop codons are shown in bold italics. Lower case letters are used for base pairs that differ from the consensus, ancestral, DNA sequence. BIME, Bacterial Interspersed Mosaic Element; *E.co.*, *E. coli*; *S.ty.*, *S. typhimurium*; *E.ae.*, *E. aerogenes*; *E.cl.*, *E. cloacae*; ancestor—putative ancestral sequence.

anchor helix F to helix C (of the same subunit). In addition, the hydrophobic character of these residues may help to stabilize the L-tryptophan binding pocket (49).

We note that, the amino acid sequence differences observed among the four Enterobacterial repressors may affect the interaction of Trp repressor with other proteins, such as WrbA, which stabilizes the repressor-operator complex in *E. coli* and *S. typhimurium* during stationary phase (63). It is not known whether the Enterobacter species make WrbA.

Alignment of the amino acid sequences of the predicted products of these four *trpR* genes provides important information to guide structure and function studies. Residues that show the least conservation (31, 32, 106, 107, and 108) are good targets for genetic and chemical modification. It should be possible to adduct fluorescence reporter groups, crosslinking agents, or other reporter molecules to the side chains of these residues (perhaps after site-directed changes of these residues to unique cysteines; see, for example, 64) with little effect on TrpR function.

Enterobacter *trpR* genes lack 3' BIME sequences

The *E. coli trpR* structural gene has an insertion of 157 base pairs located at its 3' end that includes a Bacterial Interspersed Mosaic Element (BIME, 65, 66) composed of three REP sequences (Repetitive Extragenic Palindromes, 67). The *S. typhimurium* homologue has an internal 85 base pair deletion within the BIME (29). These interspersed repeats are not found with the *E. aerogenes* or *E. cloacae trpR* genes. If the BIME-associated sequences are removed from the *E. coli* and *S. typhimurium trpR* sequences, the four *trpR* sequences can be aligned and a consensus 'ancestral' *trpR* sequence may be determined for this region (Fig. 5).

The most primitive gene from the four species is *E. aerogenes trpR*. The three other enteric *trpR* genes have insertions in their 3' ends with respect to the ancestral sequence. *E. cloacae trpR* has a seven base pair insertion, and *E. coli* and *S. typhimurium* have larger, BIME-associated insertions. The *E. cloacae* insertion results in a shift in frame, adding an amino acid to its carboxyl-terminus. The 157 base pair insertion in the *E. coli trpR* gene generates new carboxyl-termini for both *trpR* and the convergent, downstream URF, and includes a large palindromic sequence (15) that may act as a transcription terminator for one or both genes. The 85 base pair internal deletion within the *S. typhimurium* BIME removes the large palindrome and leaves a perfect 19 base pair direct repeat on each side of the deletion junction. The only two base pair differences between the parts of the BIME common to *E. coli* and *S. typhimurium* occur in the repeat 3' to the deletion junction (29). If the *E. aerogenes* or *E. cloacae* genomes once carried a BIME at this location, no trace of it remains. REP elements are found in Enterobacter genomes (68), however, the sequences of the Enterobacter BIMEs differ slightly from those of *E. coli* and *S. typhimurium* (69).

The *trpR* regulatory regions

Similarity to the 3' end of 16s RNA is low for each *trpR* ribosome binding site (RBS, 70, Fig. 6). The position of the RBS homology is nearly identical in each case, but the sequences differ somewhat. Each of the *trpR* genes encodes AUG as the start codon and has an AT base pair three base pairs in front of the start codon. The RBS sequences for *E. coli* and *S. typhimurium* are identical, and, like the *E. aerogenes* sequence, have a potential hairpin structure which partially overlaps the RBS and may inhibit translation initiation (Fig. 6b). The hairpin is shifted one base

closer to the RBS for *E. aerogenes*, is poorly conserved in its primary sequence, but has several compensatory changes that preserve its ability to form intra-strand base pairs, suggesting that it forms and functions *in vivo*. However, a 32 base pair deletion in *E. cloacae* removes this hairpin structure, suggesting that its function is dispensable. The DNA sequence of *E. coli trpR* in this region contains two direct repeats of the sequence 5'ACAA (Fig. 6b). Because deletions occur between direct repeats (71–73), the *E. coli trpR* sequence might resemble the ancestral sequence. Less efficient translation of *E. cloacae trpR* may result from a poorer RBS (70). If this hairpin serves as a terminator for the *slt* gene in *E. coli*, *S. typhimurium*, and *E. aerogenes*, then another sequence may perform this function for *E. cloacae*, or its *slt* transcript may extend into *trpR*.

The rate of transcription initiation of the *E. coli trpR* promoter is about one-tenth that of the strong *trp* promoter (74, 75). The intrinsic strength of the *trpR* promoter increases twofold in *dam*[−] *E. coli* strains, which do not methylate adenines in the sequence, 5' GATC (76). The *dam* sequence overlapping the −35 region of the *trpR* promoter is conserved among *E. coli*, *S. typhimurium*, and *E. cloacae* (Fig. 6a). It has been proposed that a burst of TrpR synthesis from one or both hemimethylated *trpR* promoters may occur after DNA replication in *E. coli*, coordinating *trpR* synthesis with the cell-cycle (76).

Computer analysis of the *E. coli trpR* sequence reveals two potentially strong promoters. One has the identified startpoint of transcription (15). A second is upstream, and has −35 and −10 regions shifted by 4 and 5 base pairs with respect to the first (Fig. 6a, 77). The upstream promoter is conserved between *E. cloacae* and *E. coli*, whereas the downstream promoter of *E. cloacae trpR* has a 3 base pair deletion in its −10 region, resulting in a much weaker match with consensus (Fig. 6a, 77). The deletion places the *E. cloacae trpR* promoter closer to its operator. The upstream promoter of *E. cloacae trpR* has a spacing of 18 base pairs between its −10 and −35 regions; the spacing is 17 base pairs for the *E. coli* upstream promoter. The intergenic region of *S. typhimurium* has a sequence identical to *E. coli* (29). The region upstream of the *trpR* gene including the operator and promoter sequences was not recovered in the clone of *E. aerogenes trpR*.

The *E. cloacae trpR* operator differs from the *E. coli*, and *S. typhimurium* operators by one base pair change in each half-site. The change from T to G at −3 is predicted to improve binding, and the change from C to T at +3 is predicted to weaken binding, based on genetic evidence (78). Together, these changes may compensate for one another. Thus, we predict that the *E. cloacae trpR* operator will be similar in strength to the *E. coli trpR* operator.

Enterobacter genes flanking *trpR*

The *slt* gene, encoding soluble lytic transglycosylase (79, 80), lies upstream of *trpR* and is transcribed in the same direction. The carboxyl-termini of the *E. coli*, *S. typhimurium*, and *E. cloacae* Slt proteins are conserved in amino acid sequence (data not shown). An unidentified reading frame (URF) of yet unknown function is present downstream of *E. coli trpR* (Bogosian, unpublished). DNA sequence alignment reveals the characteristic pattern of third position silent changes (81) confirming that this URF is transcribed in a direction opposite to *trpR* (counterclockwise on the *E. coli* map). Coding sequences for the carboxyl-termini of *trpR* and the URF overlap by eight base pairs in *E. aerogenes* and by four base pairs in *E. cloacae*. In contrast,

BIME associated insertions separate these coding sequences by 152 base pairs in *E.coli* and by 67 base pairs in *S.typhimurium* (Fig. 5).

Phylogeny

Table 2 summarizes the nucleotide and amino acid differences among the four trpR sequences. Most of the nucleotide differences (from 50 to 70%) occur in the third positions of codons and have no consequence for the amino acid sequences of the proteins.

This phenomenon has been observed in many sequence comparisons among closely related bacterial genera and probably reflects the freedom of the nucleotide sequence to diverge without constraints other than those imposed by functional requirements of the protein product and the genetic code (81). Nucleotide differences in the first positions of codons outnumber nucleotide differences in the second positions by a ratio that ranges between 1.4 to 1 and 7 to 1. This excess of first position changes over second position changes has been observed in sequence

Figure 6. (a) Nucleotide sequences of the *E. coli*, *S. typhimurium*, *E. aerogenes*, and *E. cloacae* *trpR* 5' flanking region. The intergenic region between the presumed termination codon (*tga*), of *slt* (80) and the initiation codon (*atg*) of *trpR* (15, 47) is shown. The complete nucleotide sequence of the *E. coli* Intergenic region is shown and only the differences in the other sequences are given below. The sequence of *S. typhimurium* *trpR* is identical to *E. coli* in this region (29) and is thus not shown. The presumed ribosome binding sites (RBS) are shown in uppercase-bold-italics near the ends of the sequences. The experimentally determined transcription start-site for *E. coli* *trpR* is indicated by a bullet and is numbered +1 (15). The operator sequences are shown in uppercase italics. The *trpR* operators lie across the presumed transcription start-sites. The experimentally determined set of -10 and -35 hexamers of the *E. coli* *trpR* promoter are indicated in uppercase bold typeface with dashed overlines. The set of -10 and -35 hexamers for *E. coli* *trpR* inferred on the basis of a promoter homology index algorithm (77) are indicated by uppercase with dashed overlines. The underlined GATC sequence at -35 is a substrate for *dam* methylase of *E. coli*. Gaps in the alignment are indicated by asterisks. Part of the sequence of *E. aerogenes* was not cloned, indicated by the double lines. *E. co.*, *E. coli*; *S. ty.*, *S. typhimurium*; *E. ae.*, *E. aerogenes*; *E. cl.*, *E. cloacae*. **(b)** Predicted mRNA's with possible secondary structures. Compensatory changes in the presumed *E. aerogenes* hairpin are indicated in uppercase bold typeface. Asterisks show the start and end points of the 32 base pair gap in the *E. cloacae* sequence. A four base pair direct repeat is underlined in the *E. coli* sequence. Other symbols are as in (a).

Table 2. Nucleotide and amino acid changes in *trpR*

	Comparison ¹					
	<i>E.co./S.ty.</i>	<i>E.co./E.ae.</i>	<i>E.co./E.cl.</i>	<i>S.ty./E.ae.</i>	<i>S.ty./E.cl.</i>	<i>E.ae./E.cl.</i>
Nucleotide changes ²						
1st position	3 (38)	20 (27)	24 (30)	23 (29)	27 (32)	11 (17)
2nd position	1 (12)	6 (8)	3 (4)	7 (9)	4 (5)	8 (13)
3rd position	4 (50)	49 (65)	53 (66)	48 (62)	53 (63)	44 (70)
Total	8	75	80	78	84	63
% Divergence	2.4	23	24	24	26	19
Amino acid changes						
Total	3	15	18	18	21	13
% Divergence	2.7	14	17	17	19	12

¹*E. coli*, *E. coli*: *S. ty.*, *S. typhimurium*; *E. ae.*, *E. aerogenes*; *E. cl.*, *E. cloacae*.

²Numbers in parentheses are the percentage of the total changes.

comparisons of other Enterobacterial genes (82, 83), and probably reflects the fact that first position substitutions are more likely to result in conservative amino acid replacements. The number of amino acid differences observed between each pair of TrpR proteins is shown in Table 2. The TrpR proteins from *E. coli* and *S. typhimurium*, with only three amino acid differences, are the most similar pair. The TrpR proteins of *E. aerogenes* and *E. cloacae* also cluster. The phylogenetic relationships revealed by the TrpR amino acid sequence alignment agree with the results found by the alignment of GATase amino acid sequences (83) and with the phylogenetic relationships deduced from 16S rRNA cataloging data supplemented with biochemical character states based on aromatic amino acid biosynthesis (84). In turn, these results show that the Trp repressor sequence is a useful benchmark for measuring the phylogenetic distance between different enteric bacteria.

ACKNOWLEDGEMENTS

We thank I. Skrypka and R. Somerville for kindly providing bacterial strains and for providing us with the sequence of *S. typhimurium trpR* before publication. We thank R. Somerville, B. Nichols, P. Hartzell, and M. McClelland for their critical comments on the manuscript. This work was supported by National Institutes of Health Grants GM34150 (to P.Y.) and GM12629 (to D.N.A.). C.L.L. is supported by the US Department of Energy Office of Health and Environmental Research.

REFERENCES

- Yanofsky, C. (1971) *Jama* **218**, 1026–1035.
- Yanofsky, C. and Crawford, I.P. (1987) In Neidhardt, F.C., Ingraham, J.L., Low, K.B., Magasanik, B., Schaechter, M. and Umberger, H.E. (eds.), *Escherichia coli and Salmonella typhimurium*; Cellular and Molecular Biology, American Society for Microbiology, Washington, DC, Vol. 2, pp. 1453–1472.
- Somerville, R. (1992) *Prog. Nucleic Acid Res. Mol. Biol.* **42**, 1–38.
- Luisi, B.F. and Sigler, P.B. (1990) *Biochim. Biophys. Acta* **1048**, 113–126.
- Zhao, D., Arrowsmith, C.H., Jia, X. and Jardetzky, O. (1993) *J. Mol. Biol.* **229**, 735–746.
- Yanofsky, C. (1984) *Mol. Biol. Evol.* **1**, 143–161.
- Crawford, I.P. (1975) *Bacteriol. Rev.* **39**, 87–120.
- Bennett, G.N. and Yanofsky, C. (1978) *J. Mol. Biol.* **121**, 179–192.
- Grove, C.L. and Gunsalus, R.P. (1987) *J. Bacteriol.* **169**, 2158–2164.
- Zurawski, G., Gunsalus, R.P., Brown, K.D. and Yanofsky, C. (1981) *J. Mol. Biol.* **145**, 47–73.
- Heatwole, V.M. and Somerville, R.L. (1991) *J. Bacteriol.* **173**, 108–115.
- Heatwole, V.M. and Somerville, R.L. (1991) *J. Bacteriol.* **173**, 3601–3604.
- Sarsero, J.P., Wooley, P.J. and Pittard, A.J. (1991) *J. Bacteriol.* **173**, 4133–4143.
- Heatwole, V.M. and Somerville, R.L. (1992) *J. Bacteriol.* **174**, 331–335.
- Gunsalus, R.P. and Yanofsky, C. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7117–7121.
- Eisenstein, R.B. and Yanofsky, C. (1962) *J. Bacteriol.* **83**, 193–204.
- Manson, M.D. and Yanofsky, C. (1976) *J. Bacteriol.* **126**, 679–689.
- Somerville, R.L. (1966) *Science* **154**, 1585–1587.
- Bennett, G.N., Schweingruber, M.E., Brown, K.D., Squires, C. and Yanofsky, C. (1978) *J. Mol. Biol.* **121**, 113–137.
- Miozzari, G.F. and Yanofsky, C. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 5580–5584.
- Bennett, G.N., Brown, K.D. and Yanofsky, C. (1978) *J. Mol. Biol.* **121**, 139–152.
- Blumenberg, M. and Yanofsky, C. (1982) *J. Bacteriol.* **152**, 49–56.
- Blumenberg, M. and Yanofsky, C. (1982) *J. Bacteriol.* **152**, 57–62.
- Kuroda, M.I. and Yanofsky, C. (1984) *J. Bacteriol.* **159**, 1063–1064.
- Miozzari, G.F. and Yanofsky, C. (1978) *Nature* **276**, 684–689.
- Kumamoto, A.K., Miller, W.G. and Gunsalus, R.P. (1987) *Genes. Dev.* **1**, 556–564.
- Elledge, S.J. and Davis, R.W. (1989) *Genes. Dev.* **3**, 185–197.
- Arvidson, C.G. (1990) *In Vivo Regulation of the trpEDCBA, aroH, and trpR Operons of Escherichia coli by the Trp repressor* [Ph.D. thesis], University of California, Los Angeles.
- Skrypka, I. and Somerville, R.L. (1994) *DNA Sequence*, in press.
- Zagursky, R.J. and Berman, M.L. (1984) *Gene* **27**, 183–191.
- Spindler, K.R., Rosser, D.S. and Berk, A.J. (1984) *J. Virol.* **49**, 132–141.
- Chang, A.C.Y. and Cohen, S.N. (1978) *J. Bacteriol.* **134**, 1141–1156.
- Simons, R.W., Houtman, F. and Kleckner, N. (1987) *Gene* **53**, 85–96.
- Miller, J.H. (1972) *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor.
- Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
- Segal, E., Hagblom, P., Seifert, H.S. and So, M. (1986) *Proc. Natl. Acad. Sci. U. S. A.* **83**, 2177–2181.
- Kunkel, D. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 488–492.
- Cohen, G.N. and Jacob, F. (1959) *C. R. Acad. Sci. Ser. D* **248**, 3490–3492.
- Johnson, D.I. and Somerville, R.L. (1983) *J. Bacteriol.* **155**, 49–55.
- Umberger, H.E., Umberger, M.A. and Siu, P.M.L. (1963) *J. Bacteriol.* **85**, 1431–1439.
- Marmur, J. (1961) *J. Mol. Biol.* **3**, 208–218.
- Feinberg, A.P. and Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13.
- Humbert, R., Brusilow, W.S., Gunsalus, R.P., Klionsky, D.J. and Simoni, R.D. (1983) *J. Bacteriol.* **153**, 416–422.
- Clark, A.J. and Margulies, A.D. (1965) *Proc. Natl. Acad. Sci. U. S. A.* **53**, 451–459.
- Maloy, S. and Youderian, P. (1994) *Meth. Mol. Biol.* **2**, in press.
- Sanger, F., Nicklen, S. and Coulson, A. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5476.
- Singleton, C.K., Roeder, W.D., Bogosian, G., Somerville, R.L. and Weith, H.L. (1980) *Nucleic Acids Res.* **8**, 1551–1560.
- Schevitz, R.W., Otwinowski, Z., Joachimiak, A., Lawson, C.L. and Sigler, P.B. (1985) *Nature* **317**, 782–786.
- Zhang, R.G., Joachimiak, A., Lawson, C.L., Schevitz, R.W., Otwinowski, Z. and Sigler, P.B. (1987) *Nature* **327**, 591–597.
- Lawson, C.L., Zhang, R.G., Schevitz, R.W., Otwinowski, Z., Joachimiak, A.J. and Sigler, P.B. (1988) *Proteins* **3**, 18–31.
- Czaplicki, J., Arrowsmith, C. and Jardetzky, O. (1991) *J. Biomol. NMR* **1**, 349–361.
- Arrowsmith, C., Pachter, R., Altman, R. and Jardetzky, O. (1991) *Eur. J. Biochem.* **202**, 53–66.
- Zhang, H., Zhao, D., Arrowsmith, C.H. and Jardetzky, O. (1994) *J. Mol. Biol.*, submitted.
- Otwinowski, Z., Schevitz, R.W., Zhang, R.-G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F. and Sigler, P.B. (1988) *Nature* **335**, 321–329.
- Lynch, B.A. and Koshland, D.E., Jr. (1992) *FEBS. Lett.* **307**, 3–9.
- Dodd, I.B. and Egan, J.B. (1990) *Nucleic Acids Res.* **18**, 5019–5026.
- Arvidson, D.N., Shapiro, M. and Youderian, P. (1991) *Genetics* **128**, 29–35.
- Evans, J.N., Arvidson, D.N., Gunsalus, R.P. and Roberts, M.F. (1992) *Biochim. Biophys. Acta* **1160**, 156–162.
- Marmorstein, R.Q., Joachimiak, A., Sprinz, M. and Sigler, P. (1987) *J. Mol. Chem.* **262**, 4922–4927.
- Somerville, R.L. (1983) In Herrmann, K.M. and Somerville, R.L. (eds.), *Amino Acids: Biosynthesis and Genetic Regulation*, Addison-Wesley Publishing Co., Reading, MA, pp. 351–378.
- Lawson, C.L. and Carey, J. (1993) *Nature* **361**, 178–182.
- Hirel, P.H., Schmitter, M.J., Dessen, P., Fayat, G. and Blanquet, S. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 8247–8251.
- Yang, W., Ni, L. and Somerville, R.L. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5796–5800.
- Sutton, C.L., Mazumder, A., Chen, C.H. and Sigman, D.S. (1993) *Biochemistry* **32**, 4225–4230.
- Gilson, E., Saurin, W., Perrin, D., Bachellier, S. and Hofnung, M. (1991) *Nucleic Acids Res.* **19**, 1375–1383.
- Gilson, E., Saurin, W., Perrin, D., Bachellier, S. and Hofnung, M. (1991) *Res. Microbiol.* **142**, 217–222.
- Higgins, C.F., Ames, G.F.-L., Barnes, W.M., Clement, J.M. and Hofnung, M. (1982) *Nature* **298**, 760–762.
- Dimri, G.P., Rudd, K.E., Morgan, M.K., Bayat, H. and Ames, G.F. (1992) *J. Bacteriol.* **174**, 4583–4593.
- Bachellier, S., Perrin, D., Hofnung, M. and Gilson, E. (1993) *Mol. Microbiol.* **7**, 537–544.

70. Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G.D. and Gold, L. (1992) *Mol. Microbiol.* **6**, 1219–1229.
71. Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. and Inouye, M. (1966) *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77–84.
72. Farabaugh, P.J., Schmeissner, U., Hofer, M. and Miller, J. (1978) *J. Mol. Biol.* **126**, 847–863.
73. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C.O., Spritz, R.A., DeRiel, J.K., Forget, B., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Barralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) *Cell* **21**, 653–668.
74. Bogosian, G. and Somerville, R.L. (1984) *Mol. Gen. Genet.* **193**, 110–118.
75. Cho, K.O. and Yanofsky, C. (1988) *J. Mol. Biol.* **204**, 41–50.
76. Marinus, M.G. (1985) *Mol. Gen. Genet.* **200**, 185–186.
77. Harley, C.B. and Reynolds, R.P. (1987) *Nucleic Acids Res.* **15**, 2343–2361.
78. Bass, S., Sugiono, P., Arvidson, D.N., Gunsalus, R.P. and Youderian, P. (1987) *Genes. Dev.* **1**, 565–572.
79. Betzner, A.S. and Keck, W. (1989) *Mol. Gen. Genet.* **219**, 489–491.
80. Engel, H., Kazemier, B. and Keck, W. (1991) *J. Bacteriol.* **173**, 6773–6782.
81. Crawford, I.P., Nichols, B.P. and Yanofsky, C. (1980) *J. Mol. Biol.* **142**, 489–502.
82. Kaplan, J.B. and Nichols, B.P. (1983) *J. Mol. Biol.* **168**, 451–468.
83. Kaplan, J.B., Merkel, W.K. and Nichols, B.P. (1985) *J. Mol. Biol.* **183**, 327–340.
84. Ahmad, S., Weisburg, W.G. and Jensen, R.A. (1990) *J. Bacteriol.* **172**, 1051–1061.
85. Amann, E., Brosius, J. and Ptashne, M. (1983) *Gene* **25**, 167–178.
86. Miller, J.H., Ganem, D., Lu, P. and Schmitz, A. (1977) *J. Mol. Biol.* **109**, 275–301.
87. Vogel, H.J. and Bonner, D.M. (1956) *J. Biol. Chem.* **218**, 97–106.
88. Kraulis, P.J. (1991) *Appl. Crystallogr.* **24**, 946–950.