DNA sequence of the *E. coli* *trpR* gene and prediction of the amino acid sequence of Trp repressor

C.K.Singleton, W.D.Roeder*, Gregg Bogosian, R.L.Somerville and H.L.Weith

Department of Biochemistry, Purdue University, West Lafayette, IN 47907, USA

Received 10 March 1980

ABSTRACT

A DNA sequence of 1041 base pairs from a *Bam*HI fragment containing the *E. coli* *trpR* gene has been determined. With this sequence and other experimental evidence, the primary structure (88 amino acids) of the Trp repressor can be predicted. Additional features of the DNA sequence include a 22 base pair region upstream from the proposed structural gene which exhibits striking homology with the *trp* operator, thus implying that expression of the *trpR* gene may be under autogenous regulation.

INTRODUCTION

The *trpR* gene¹ codes for a repressor protein involved in control of the tryptophan biosynthetic system.² The kinetic parameters of Trp repressor-operator interaction *in vitro* have been determined³, as has the structure of *trp* operator DNA.⁴ However, the Trp repressor protein has not been obtained in purified form, nor has the structure of the *trpR* gene been determined. Recently, DNA containing a functional *trpR* gene from *E. coli* has been cloned.⁵ We now report the nucleotide sequence of this DNA and predict the primary structure of Trp repressor protein.

THE DNA SEQUENCE

Figure 1 summarizes the origin of and sequencing strategy carried out on a 1250 base pair *Bam*HI fragment from *E. coli* DNA containing the tryptophan repressor gene. A major portion of the sequence was obtained by cleaving the 1250 base pair fragment with either *Hae*III or *Hin*FI restriction endonucleases. In each case the restriction fragments were labelled at the 5' ends, digested with a second restriction endonuclease, and purified by acrylamide gel electrophoresis before sequencing. Additional DNA fragments generated with other restriction enzymes were also analyzed in order to fill in gaps, read across the *Hae*III and *Hin*FI sites, and determine complementary nucleotide sequences. In all, a sequence of 1041 base pairs, from the left *Bam*HI site to the far right *Hin*FI site, was obtained (Figure 2). At least 85% of the residues in each strand were independently determined.

Examination of the three reading frames of each strand indicate the possible coding of numerous polypeptides. All but one of these can be eliminated as the coding sequence for Trp repressor by other available experimental evidence. In vitro transcription and translation of the 1250 base pair fragment produces a protein of approximately 10,000 daltons, as determined by SDS-acrylamide gel electrophoresis (Figure 3). This is consistent with the size of only three of the possible polypeptides coded for in this region of *E. coli* DNA (Figure 2).

Figure 2. (Facing page) Nucleotide sequence of 1041 base pairs from the BamHI fragment. Only those residues for which the sequence was directly determined are shown. Cytidine residues in lower case represent modified bases which appeared as gaps on the sequencing gels. These were confirmed (except for position 120) by sequencing the complementary strand. The three possible coding regions which predict polypeptides of approximately 10,000 daltons are enclosed with $\nabla \Delta \circ$, signifying initiation points, and $\nabla \blacktriangle \bullet$, which delineate termination sites. Downward and upward arrows show the single SalI cleavage site. The G/C-rich symmetry region discussed in the text is underscored.

5' GATCCGGAACGAATATCAACATTGGCACCAGTTACCTGCAATATGTTTA 50
 5' TCAGCAGTTTGGCAATAATCGTATTTTCTCCTCAGCAGCTTATAACGCCG 100
 5' GACTAGGGGCGGGTGCGAACcTGGCTTGGCAACAGCGCCGGGCGTATCGAC 150
 5' GCAGTGGCATTGTGTCGAGAGTATTCCATTCTCCGAGACGCGCGGTTATGT 200
 3' AAACAGCTCTCATAAGGTAAGAGGCTCTGCGCGCCAATACA
 5' GAAGAACGTGCTGGCTTATGACGCTTACTACCGCTATTTTCATGGGGGATA 250
 3' CTTCTTGACGACCGAATACTGCGAATGATGGCGATAAAGTACCCCTAT
 5' AACCGACGTTGATGAGCGCCACGG GATCCGCA 300
 3' TTGGCTGCAACTACTCGCGGTGCCCTTACCCCTGCAGCAATGACTAGGCGT
 5' CGTTTATGATATGCTATCGTACTCTTTAGCGAGTACAACCGGGGGAGGCA 350
 3' GCAAATACTATACGATAGCATGAGAAATCGCTCATGTTGGCCCCCTCCGT
 5' TTTTGCTTCCCCCGCTAACAATGGCGACATATTATGGCCCAACAATCACC 400
 3' AAAACGAAGGGGGCGATTGTTACCGCTGTATAATACCGGGTTGTTAGTGG
 5' CTATTGACGAGCGATGGCAGAACAGCGTCACCAGGAGTGGTTACGTTTTG 450
 3' GATAAGTCGTCGCTACCGTCTTGTGCGCAGTGGTcCTCACC AATGCAAAAC
 5' TCGACCTGCTTAAGAATGCCTACCAAAACGATCTCCATTTACCGTTGTTA 500
 3' AGCTGGACGAATTCCTACGGATGGTTTTGCTAGAGGTAAATGGCAACAAT
 5' AACCTGATGCTGACGCCAGATGAGCGCGAAGCGTTGGGGACTCGCGTGCG 550
 3' TTGGACTACGACTGCGGTCTACTCGCGCTTCGCAACCCCTGAGCGCACGC
 5' TATTGTCGAAGAGCTGTTGCGCGGCGAAATGAGCCAGCGTGAGTTAAAAA 600
 3' ATAACAGCTTCTCGACAACGCGCCGCTTTACTCGGTGCGCACTCAATTTTT
 5' TGAACCTGGCGCAGGCATCGCGACGATTACGCGTGGATCTAACAGCCTGA 650
 3' ACTTGAGCCGCGTCCGTAGCGCTGCTAATGCGCACCTAGATTGTCGGACT
 5' AAGCCGCGCCCGTCTGAGCTGCGCCAGTGGCTGGAAGAGGTGTTGCTGAAA 700
 3' TTCGGCGCGGGCAGCTCGACGCGGTACCGACCTTCTCCACAACGACTTT
 5' AGCGATTGATTTTGTAGGCCTGATAAGACGTGGCGCATCAGGCATCGTGC 750
 3' TCGCTAACTAAACATCCGGACTATTCTGCACCGCGTAGTCCGTAGCAGC
 5' ACCGAATGCCGGATGCGGCGTGAACGCCTTATCCGTCTACAAATACCCG 800
 3' TGGCTTACGGCCTACGCCGCACTTGCGGAATAGGCAGGATGTTTATGGGC
 5' TAATTTCAATATGTTTGGTAGGCATGATAAGACGCGGCAGCGTCGCATCA 850
 3' ATTAAGTTATACAAACCATCCGTA_ΔCTATTCTGCGCCGTCGCAGCGTAGT
 5' GGCCTTAATACACGGCATTATGAAACGGACTCAGCGCCAGGATCACCGC 900
 3' CCGCGAATTATGTGCCGTAATACTTTGCCTGAGTCGCGGTCTAGTGGCG
 5' cTGGTGATAGACGTGGCGCGAGTGAGTTTCCGGCGGTAAACACGCCGA 950
 3' GACCACTATCTGCGACCGCGCTCACTCAAAGGGCC
 3' AGCGGGGAAGGAATGCTGGCTAGAGTAGTTATGGCCATATTGCGCTGTAG 1000
 3' TGCCCTGGTTTCGCGGAGTGGAAGTGCGTGAAAAAGGTCTCA 1041

FIGURE 2

Figure 3. *In vitro* transcription and translation of *trpR* DNA. 5 or 10 micrograms of the 1.25 kb *Bam*HI fragment was added to a standard S-30 system.⁶ (lanes 1 and 2 respectively). Lane 3 contained pBR322 DNA, while lane 4 represents no added DNA. Incubation was for 60 minutes, followed by electrophoretic separation of the ³⁵S-Met-labelled products on 20% SDS-polyacrylamide gels. Autoradiography was done at -70° C for 24 hours. The migration positions of proteins of known molecular weight are indicated.

cleaved asymmetrically with restriction endonuclease *Sal*I and the two resulting fragments cloned separately, a *trpR* negative phenotype is obtained in each case.⁵ This is significant in that two of the three possible coding sequences, both present in the leftward strand, terminate outside of the *Sal*I site (Figure 2). However, the *Sal*I site lies within the initial portion of the third coding sequence located on the rightward strand. Similarly, when a two or three base pair addition is made within the *Sal*I site, a *trpR* negative phenotype is observed on cloning the 1250+(2,3) base pair fragment, further emphasizing the importance of the integrity of the *Sal*I site for *trpR* expression (G. Bogosian, R. Somerville, unpublished results). Finally, RNA polymerase binding studies on the 1250 *Bam*HI fragment have demonstrated the existence of a single binding site. This site is contained within the *Alu*I-*Sal*I fragment from position 88 to 452 (S. Kuhstoss, R. Somerville, unpublished results). The localization of a polymerase binding site within this region coupled with the requirement for an intact *Sal*I site for functional repressor is consistent only with the assignment of the DNA region shown in Figure 4 as being the *trpR* coding sequence.

Based on these considerations, we conclude that the polypeptide shown in Figure 4 is likely to be the Trp repressor. The repressor monomer probably consists of 88 amino acids (MW 10,250 Daltons). However, thirty nucleotides to the right of the indicated start codon there is another methionine codon, and we cannot rule out the possibility that the repressor may be only 78 amino acids in length. The second ATG, however, is considered unlikely as the site of initiation when the region immediately upstream from these methionine codons is compared to initiation sequences compiled by Atkins.⁷

```

MetAlaGlnGln SerProTyrSer AlaAlaMetAla GluGlnArgHis GlnGluTrpLeu
ATGCCCAACAA TCACCCTATTCA GCAGCGATGGCA GAACAGCGTCAC CAGGAGTGGTTA
      395           407           419           431           443
ArgPheValAsp LeuLeuLysAsn AlaTyrGlnAsn AspLeuHisLeu ProLeuLeuAsn
CGTTTTGTCGAC CTGCTTAAGAAT GCCTACCAAAAC GATCTCCATTTA CCGTTGTAAAC
      455           467           479           491           503
LeuMetLeuThr ProAspGluArg GluAlaLeuGly ThrArgValArg IleValGluGlu
CTGATGCTGACG CCAGATGAGCGC GAAGCGTTGGGG ACTCGCGTGCCT ATTGTCAAGAG
      515           527           539           551           563
LeuLeuArgGly GluMetSerGln ArgGluLeuLys MetAsnSerAla GlnAlaSerArg
CTGTTGCGCGGC GAAATGAGCCAG CGTGAGTTAAAA ATGAACTCGGCG CAGGCATCGCGA
      575           587           599           611           623
ArgLeuArgVal AspLeuThrAla Opa1
CGATTACGCGTG GATCTAACAGCC TGA
      635           647

```

Figure 4. The coding sequence for the *trpR* gene beginning at position 384 and ending at position 650 is shown. The amino acid sequence of the *trpR* protein is presented above the nucleotide sequence.

There is a TAA termination codon 18 base pairs upstream from the first methionine codon, but there is neither a TAA nor a TGA within 48 base pairs upstream from the second methionine. This is consistent with Atkins' observations only if the first ATG is the actual site of initiation. Previous work from other laboratories has shown that active Trp repressor protein has a molecular weight of 58,000 to 60,000 Daltons, as determined by gel filtration.⁸⁻¹⁰ Thus, the mature Trp repressor may exist as a hexamer of identical subunits, or the polypeptide predicted here may form a complex with other proteins *in vivo*.

The relatively small size of the Trp repressor monomer is consistent with the known properties of other repressors. The *cro* proteins of phages lambda and 434 are repressors of genes expressed early in phage development. These regulatory proteins contain 66 and 71 amino acids, respectively.¹¹⁻¹³ Lambda repressor, the product of the *cI* gene, contains 236 amino acids, yet the first 92 residues comprising the amino terminal domain are sufficient for operator recognition.^{14,15} Finally, recent studies on *lac* repressor-operator interaction have shown that residues 1-51 of *lac* repressor make all the operator DNA contacts characteristic of intact *lac* repressor, which is 347 amino acids in length.¹⁶

Features of the DNA Sequence

Analysis of codon usage in the *trpR* gene reveals several interesting features. First, there appears to be asymmetric selection for so-called

"weak" and "strong" codons, similar to that observed in the lacI gene by Fiers and Grosjean.¹⁷ Specifically, these authors have pointed out that in the lacI gene, from which lac repressor is made in low levels, T's are preferred over C's as the third letter in "weak" codons (A/T-A/T-Y, where Y=T or C), and C's are preferred over T's in "strong" codons (G/C-G/C-Y).¹⁸ They further noted that this preference is exactly opposite to that found for bacteriophage MS2 coat protein gene, a protein made in very high levels in phage infected cells. For the trpR gene preference is shown (Table I) for T in two of the four "weak" codons (Phe, Ile), and absolute preference is shown for C with "strong" codons (Pro, Ala, Gly). Considering the seven amino acids for which this codon selection can be made, the trpR gene shows weak or strong codon usage 69% of the time. Perhaps such codon selectivity is significant in light of the very low levels of protein produced by the trpR and lacI genes.

Another feature of codon usage in the trpR gene concerns the marked asymmetric occurrence of the four nucleotides in the third position of codons. For example, T residues occur in the third position only 15.9% of the time. This seems unusual considering the fact that 65 out of 88 codons used in the trpR gene could have T in the third position by degeneracy of the code. Furthermore, the percentage of G residues in the third position is relatively high at 36.4%, yet the number of codons which could possibly use G in the third position is only 3 more than those which could employ T.

One additional feature of the DNA sequence must be noted. Inspection

Table 1. Codon Usage in the trpR gene of E. coli

	U	C	A	G	
U	Phe — 1 0 Leu — 5 3	Ser — 0 0 2 2	Tyr — 1 0c — 0 Am — 0	Cys — 0 0 op — 1 Trp — 1	U C A G
C	Leu — 1 1 4	Pro — 0 1 1	His — 1 Gln — 3 4	Arg — 4 4 2 0	U C A G
A	Ile — 1 0 Met — 5	Thr — 0 1 1	Asn — 1 3 Lys — 1	Ser — 0 1 Arg — 0 0	U C A G
G	Val — 0 2 0 2	Ala — 0 3 3 3	Asp — 3 1 Glu — 4 4	Gly — 0 1 0 1	U C A G

Figure 5. Comparison of the DNA sequence preceding the *trpR* gene (residues 313-334) with the *trp* operator and an analogous Trp repressor target sequence lying within the *aroH* promoter. Homologous regions are boxed in.

of the DNA sequence in the region upstream from the *trpR* gene reveals the presence of a 22 base pair hyphenated palindrome possessing a striking homology with two target sequences for the Trp repressor protein (Figure 5). This sequence is 73% homologous with the *trp* operator⁴ and 68% homologous with a target site upstream from *aroH*,¹⁹ another gene whose expression is regulated by Trp repressor. This finding suggests that the expression of the *trpR* gene is autogenously regulated and provides a rationale for an earlier finding that there is a lack of proportionality between the *trpR* dosage and the repression of the *trp* operon.⁵

Finally, immediately to the right of this 22 base pair region exists an extensive, G/C-rich, hyphenated palindrome (Figure 2). Hairpin formation in this region could produce a stem of 13 base pairs with two mismatches and a loop of either three or five bases. Whether or not this region is functionally important in *trpR* gene expression remains to be determined.

CONCLUSION

The Trp repressor protein exists in very low levels within the *E. coli* cell, and thus has not been amenable to direct protein sequence analysis. The present work, however, demonstrates that the DNA sequence of the *trpR* gene region in conjunction with *in vitro* transcription/translation data, RNA polymerase binding studies, and a requirement for an intact *SalI* site for *trpR* expression allows for the prediction of the Trp repressor amino acid sequence. To obtain more direct evidence that the predicted structure for Trp repressor is correct, the *trpR* gene is currently being combined with a high efficiency promoter/ribosome binding sequence²⁰ in order to obtain plasmid derivatives which will stimulate high level production of Trp repressor for direct amino acid composition and sequence analysis.

MATERIALS AND METHODS

The following materials were obtained from commercial sources:

restriction endonucleases (New England Biolabs, Inc., Bethesda Research Laboratories, Inc.), T4 polynucleotide kinase and calf intestinal alkaline phosphatase (Boehringer Mannheim), and [γ - 32 P]-ATP (specific activity > 2000 Ci/mmol, ICN).

Cloning of the *E. coli* trpR gene was carried out as follows: DNA from the N-defective specialized transducing phage serB⁺trpR⁺59⁵ was purified by equilibrium banding in a CsCl density gradient in the presence of ethidium bromide. The purified DNA was digested with restriction endonuclease BamHI and ligated with BamHI-cleaved pBR322 DNA at 16.5°C for 24 hr. Ligation mixtures were used directly to transform the *E. coli* serB-trpR deletion strain 122-1.⁵ Among the ampicillin-resistant tetracycline-sensitive transformants, several were found which exhibited sensitivity to 5-methyltryptophan, a phenotype which is diagnostic for the presence of a functional trpR gene. Plasmids prepared from these isolates invariably contained a new 1.25 kb BamHI fragment inserted into the BamHI site of the plasmid in either of the two possible orientations. Anthranilate synthase activity was shown to be reduced 10 to 20 fold in cells of strain 122-1, containing plasmid pBR322 which possessed the 1.25 kb BamHI fragment in either orientation, when grown in the presence of tryptophan.⁵

The 1250 base pair DNA fragment was isolated preparatively by digestion of plasmid ptrpR3⁵ with BamHI. The digestion mixture was applied to a 5% polyacrylamide slab gel (0.6 x 20 x 20 cm) and run overnight at 150 volts. The DNA bands were located by fluorescence quenching using a uv transilluminator and fluorescent indicator TLC plates. The DNA was eluted from the gel by the crush and soak method.²¹

Ten to twenty picomoles of the 1250 base pair BamHI fragment were used as starting material for nearly all sequencing runs. This material was digested with an appropriate restriction endonuclease, and either the total digest was end-labelled or individual unlabelled fragments were separated and isolated as described above. One exception to this was the direct isolation of a specific fragment from MspI digested plasmid ptrpR3.

Sequencing was carried out on 5'-end labelled restriction fragments according to the method of Maxam and Gilbert.²¹ DNA fragments were dephosphorylated by dissolving in 45 μ l of 0.05 M Tris-HCl, pH 8.8, 0.005 M MgCl₂, 0.1 mM ZnCl₂, containing 0.25 units of calf intestinal phosphatase and incubated 30 minutes at 37°. The reaction was terminated by adding 5 μ l of 0.1 M EDTA·2Na and heating for 10 minutes at 70°. These conditions were shown to irreversibly destroy the phosphatase activity. The fragments were

then ethanol precipitated and 5'-end labelling was carried out under conditions similar to those of Maxam and Gilbert.²¹ Single strand DNA fragments or duplex DNA fragments labelled at only one end were obtained either by strand separation²¹ or by secondary restriction enzyme digestion followed by separation on 5% polyacrylamide gels. The G, A>G, T+C, and C reactions²¹ were run on each fragment. The four reactions were divided into aliquots, and 20%, 10%, or 8% polyacrylamide gels (0.15 X 20 X 40 cm) were used to separate the polynucleotide chain series generated in each reaction.

ACKNOWLEDGEMENTS

We would like to thank Pekka Mantsala for carrying out the *in vitro* transcription/translation experiments, and Irwin Tessman for conducting computer analysis of the sequence data. This work was supported by NIH Grants GM 22131 and CA 23332. CKS and WDR were supported by NIH training grant 5T32GM07076, and HLW by NIH Research Career Development Award CA 00447. This is paper No. 7842 from the Purdue University Agricultural Experiment Station.

*Present Address Basel Institute for Immunology, Basel 5, Switzerland

REFERENCES

1. Cohen, G. and Jacob, F. (1959) *Compt. Rend. Acad. Sci. Paris* **248**, 3490-3492.
2. Morse, D. E. and Yanofsky, C. (1969) *J. Mol. Biol.* **44**, 185-193.
3. Rose, J. K. and Yanofsky, C. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 3134-3138.
4. Bennett, G. N., Schweingruber, M. E., Brown, K. D., Squires, C. and Yanofsky, C. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 2351-2355.
5. Roeder, W. and Somerville, R. L. (1979) *Mol. Gen. Genet.* **176**, 361-368.
6. Zubay, G., Chambers, D. A. and Cheong, L. D. (1971) in *The Lactose Operon* (Beckwith, J. R. and Zipser, D., Eds.) pp. 375-391. Cold Spring Harbor Laboratory, New York.
7. Atkins, J. F. (1979) *Nucleic Acids Res.* **7**, 1035-1041.
8. Zubay, G., Morse, D. E., Schrenk, W. J., and Miller, J. H. M. (1972) *Proc. Natl. Acad. Sci. USA* **69**, 1100-1103.
9. Shimizu, Y., Shimizu, M., and Hayashi, M. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 1990-1994.
10. Squires, C. L., Rose, J. K., Yanofsky, C., Yang, H., and Zubay, G. (1973) *Nature New Biol.* **245**, 131-133.
11. Roberts, T. M., Shimatake, H., Brady, C. and Rosenberg, M. (1977) *Nature* **270**, 274-275.
12. Hsiang, M. W., Cole, D. C., Takeda, Y. and Echols, H. (1977) *Nature* **270**, 275-277.
13. Grosschedl, R. and Schwarz, E. (1979) *Nucleic Acids Res.* **6**, 867-881.
14. Sauer, R. T. (1978) *Nature* **276**, 301-302.
15. Sauer, R. T., Pabo, C. O., Meyer, B. J., Ptashne, M. and Backman, K. C. (1979) *Nature* **279**, 396-400.
16. Ogata, R. T. and Gilbert, W. (1979) *J. Mol. Biol.* **132**, 709-728.
17. Fiers, W. and Grosjean, H. (1979), *Nature* **277**, 328.

18. The use of "weak" codons refers to the fact that the first two letters of these codons are A or T, thus giving rise to relatively weak base pairing upon interaction with the appropriate anticodons. Similarly, "strong" codons, as used here, have a G or C as the first two letters and thus possess stronger base pairing properties.
19. Zurawski, G., Gunsalus, R., Brown, K. and Yanofsky, C., manuscript in preparation.
20. Roberts, T. M., Kacich, R. and Ptashne, M. (1979) Proc. Nat. Acad. Sci. USA 76, 760-764.
21. Maxam, A. M. and Gilbert, W. (1974) Proc. Natl. Acad. Sci. USA 74, 560-564.