DNA sequence of the *araC* regulatory gene from *Escherichia coli* B/r

C.Garrett Miyada, Arnold H.Horwitz, Laura G.Cass, Josef Timko and Gary Wilcox

Department of Microbiology and Molecular Biology Institute, University of California at Los Angeles, Los Angeles, CA 90024, USA

Received 8 September 1980

ABSTRACT

The DNA sequence of the *araC* regulatory gene from *Escherichia coli* B/r has been determined by the base-specific chemical cleavage reactions of Maxam and Gilbert. An open reading frame is found which codes for a protein of 292 amino acids. A nonsense mutation, *araC5*, is shown to result from a G to A transition at nucleotide 429 converting the tryptophan codon TGG to the amber codon TAG. A deletion which does not recombine with any known point mutation in *araC*, Δ (*araC*)719, removes all but the last 22 codons of the gene.

INTRODUCTION

The *araC* gene in *Escherichia coli* codes for a regulatory protein which controls the expression of at least six structural genes which are involved in the transport and catabolism of L-arabinose (1, 2). The *araC* protein regulates initiation of transcription of the *araBAD* operon (3) and in addition autogenously controls its own synthesis (4). The gene is transcribed in a clockwise direction on the standard *E. coli* linkage map (5). The DNA sequence of the *araC* promoter and the start site of *araC* mRNA have been determined (6). We have recently shown that *araC* mRNA contains a 150 nucleotide leader by sequencing the DNA corresponding to the promoter-proximal region of the *araC* gene (7). In order to precisely define the molecular basis by which *araC* protein controls gene expression it is necessary to determine the DNA sequence of the entire gene. In this paper we report the DNA sequence of the *araC* gene and the sequence changes resulting from a nonsense mutation and a deletion in *araC*.

MATERIALS AND METHODS

Materials

Restriction endonucleases *Taq* I and *Ava* I were purified by the methods of Sato *et al.* (8) and Murray *et al.* (9), respectively. *Hae* III was

purchased from Calbiochem-Behring Corporation, La Jolla, CA. Ava II was from Bethesda Research Laboratories, Rockville, MD. Hinf I and Sau 3A were gifts from W. Salser and M. Komaromy, respectively, both of the Molecular Biology Institute, UCLA. Bacterial alkaline phosphatase (BAP) was purchased from Worthington Biochemical Corporation, Freehold, N.J.. Polynucleotide kinase was purchased from New England Biolabs, Beverly, MA. Reverse transcriptase was obtained from the National Cancer Institute, Bethesda, MD. (γ - 32 P) ATP was prepared as described by Johnson and Walseth (10). (α - 32 P) deoxynucleotides were purchased from Amersham, Arlington Heights, IL. Cronex 4 X-Ray film and Cronex Hi Plus intensifying screens were from E.I. DuPont de Nemours, Wilmington, DE.

Plasmids

Plasmids used for sequencing the araC gene and mutations in the araC gene were pLGC10, pLGC20, pAH105, pAH25 and pCH11. Plasmids pLGC10, pLGC20, and pCH11 have been described previously (7, 11). Plasmid pAH105 was constructed by subcloning Eco RI* restriction fragments from plasmid pTB1 (12) into the Eco RI site of a modified pBR322 plasmid. Plasmid pAH105 was found among Ara⁺, Ap^r transformants in an araC deletion background. Plasmid pAH25 was constructed by in vivo recombination of the araC5 nonsense mutation onto the ara deletion 738-containing plasmid, pAH17, as described by Horwitz et al. (11). Plasmids were grown in derivatives of E. coli K-12 strain RR1 and the DNA was purified as described by Norgard et al. (13).

Isolation and end-labeling of restriction fragments

Restriction fragments used for sequencing were purified on 5% polyacrylamide or 1.5% agarose gels buffered in TBE (10.8 gm Trizma-base, 5.5 gm boric acid, and 0.93 gm Na₂EDTA per liter). The restriction fragments were visualized by U.V. shadowing (14), cut out of the gel, electroeluted out of the gel matrix, and concentrated by ethanol precipitation. DNA obtained from agarose gels was further purified after the electroelution step by passing it over a 0.6-1.2 ml DE-52 column (15). Restriction fragments were dephosphorylated using 1.4 units of bacterial alkaline phosphatase and incubating in 100-300 μ l of 50 mM Tris-HCl, pH 8.0, for one hour at 65⁰ C. The BAP was removed by successive phenol, phenol-chloroform, and ether extractions. Dephosphorylated restriction fragments were 5' end-labeled with (γ - 32 P) ATP and polynucleotide kinase essentially as described by Maxam and Gilbert (14). Restriction fragments were 3' end-labeled with (α - 32 P) dNTP and reverse transcriptase by filling in a 5' extended restriction site. The reverse transcription

reaction mixture included 5-20 pmoles of restriction fragment, 100 μ Ci of an appropriate (α - 32 P) dNTP (400 Ci/mmol), 33 μ M of any necessary cold dNTP, 6 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$, 50 mM KCl, 1 mM dithiothreitol, 10% glycerol and 50 units of reverse transcriptase, in a volume of 30 μ l. The reaction was incubated at 37 $^{\circ}$ C for at least 90 min. and terminated as the above kinase reaction. Single end-labeled fragments were usually generated by a secondary restriction cut. Alternatively, strand separation was used to create single end-labeled restriction fragments.

DNA sequence analysis

Single end-labeled restriction fragments were subjected to the base-specific chemical cleavage reactions of Maxam and Gilbert (14). The G, G+A, A+C, C+T and C reactions were used. The products were analyzed on 0.4 mm thick gels (16) using acrylamide concentrations of 15-16%, 12% and 8%. The gels were developed at -70 $^{\circ}$ C with or without intensifying screens. The final DNA sequence was analyzed by the computer program of Queen and Korn (17).

RESULTS AND DISCUSSION

Sequence analysis of the wild-type araC gene

The end-labeled restriction fragments used to determine the nucleotide sequence of the promoter-distal region of the araC gene are shown in Figure 1. As seen in the figure, the sequencing strategy employed resulted in the determination of the nucleotide sequence for both strands of DNA throughout the entire region. In addition, every restriction site used for

Figure 1. DNA sequencing strategy for the promoter-distal region of the araC gene. Pertinent restriction sites are marked by vertical bars above the heavy horizontal line. The numbers below the heavy horizontal line represent the base pair number of the gene using the araC mRNA start as 1. The horizontal arrows above the line denote sequenced restriction fragments on the upper strand (5' to 3' reading left to right); the horizontal arrows below denote sequenced restriction fragments labeled on the lower strand. Open circles at the ends of the horizontal arrows represent 3'-end labeling; closed circles represent 5'-end labeling. The arrowheads extend as far as the sequence could be determined for that fragment.

end-labeling or for a secondary restriction cut was sequenced across from a nearby restriction site. This eliminated the possibility of missing an internal sequence between two identical and adjacent restriction sites.

The entire nucleotide sequence for the araC gene is shown in Figure 2. If the methionine codon centered at position 147 is taken as the first amino acid, the araC gene codes for a protein of 292 amino acids with a molecular weight of 33,315. Two other potential methionine start codons in the same reading frame are centered at positions 127 and 138 and have been discussed in a previous paper (7). Isolated araC protein run on SDS-polyacrylamide gels was found to have a subunit molecular weight of 28,000 (18). The apparent discrepancy between the molecular weight figures may be due either to aberrant mobility of araC protein on SDS-polyacrylamide gels, proteolytic processing, or a combination of the two. The location of the Ava I site starting at position 961 within the open reading frame is consistent with the phenotype of the plasmid pLGC10. This plasmid contains the araC promoter and araC gene up to the Ava I site and is unable to complement araC⁻ alleles (7). Since the coding region of araC corresponding to the terminal 20 amino acids is not present in pLGC10, it is not surprising that no functional araC protein is produced.

DNA sequence of an amber mutant

An araC nonsense mutation, araC5 (19), was cloned onto a hybrid ara plasmid by in vivo recombination (11). The resulting recombinant plasmid, designated pAH25, was shown by the following criteria to contain the araC5 mutation. First, the plasmid failed to complement strain LA3, containing araC deletion 719, but was able to complement strain LA304 which contains araC deletion 719 and the amber suppressor mutation, supE. Second, the araC5 mutation was previously shown by genetic mapping to be located near the center of the araC gene (19). An ara point mutation on a plasmid can be readily mapped because recombination between a mutation on the plasmid and mutations on the chromosome can be detected as Ara⁺ papillae on indicator plates after two days of incubation (11). The araC5 mutation was shown to map in precisely the same location in araC on the plasmid pAH25 as it did on the chromosome by transforming this plasmid into seven strains containing different araC⁻ alleles and scoring for recombinants.

The base pair change resulting in araC5 was determined by DNA sequence analysis to be a G to A transition at position 429 in the araC gene (see Figure 3). This base pair change results in the conversion of a tryptophan codon, TGG, to the amber codon, TAG, as was expected from the genetic

```

-40      -30      -20      -10      1      10      20      30
GATTATAGACACTTTTGTGACGGCTTTTGTTCATGGCTTGGTCCCGCTTTGTTACAGAAATGCTTTTAATAAGCGGGGT
40      50      60      70      80      90      100      110
TACCGGTTGGTTAGCGAGAGAGCCAGTAAAGACGACGACGGCAATGTCTGATGCAATATGGACAATTGGTTTCT
120      130      140      150      160      170      180
TCTCTGAATGGCGGGAGTATGAAAAGT ATG GCT GAA GCG CAA AAT GAT CCC CTG CTG CCG GGA TAC
Met Ala Glu Ala Gln Asn Asp Pro Leu Leu Pro Gly Tyr
190      200      210      220      230      240
TCG TTT AAT GCC CAT CTG GTG GCG GGT TTA ACG CCG ATT GAG GCC AAC GGT TAT CTC GAT
Ser Phe Asn Ala His Leu Val Ala Gly Leu Thr Pro Ile Glu Ala Asn Gly Tyr Leu Asp
250      260      270      280      290      300
TTT TTT ATC GAC CGA CCG CTG GGA ATG AAA GGT TAT ATT CTC AAT CTC ACC ATT CGC GGT
Phe Phe Ile Asp Arg Pro Leu Gly Met Lys Gly Tyr Ile Leu Asn Leu Thr Ile Arg Gly
310      320      330      340      350      360
CAG GGG GTG GTG AAA AAT CAG GGA CGA GAA TTT GTT TGC CGA CCG GGT GAT ATT TTG CTG
Gln Gly Val Val Lys Asn Gln Gly Arg Glu Phe Val Cys Arg Pro Gly Asp Ile Leu Leu
370      380      390      400      410      420
TTC CCG CCA GGA GAG ATT CAT CAC TAC GGT CGT CAT CCG GAG GCT CGC GAA TGG TAT CAC
Phe Pro Pro Gly Glu Ile His His Tyr Gly Arg His Pro Glu Ala Arg Glu Trp Tyr His
430      440      450      460      470      480
CAG TGG GTT TAC TTT CGT CCG CGC GCC TAC TGG CAT GAA TGG CTT AAC TGG CCG TCA ATA
Gln Trp Val Tyr Phe Arg Pro Arg Ala Tyr Trp His Glu Trp Leu Asn Trp Pro Ser Ile
490      500      510      520      530      540
TTT GCC AAT ACG GGG TTC TTT CGC CCG GAT GAA GCG CAC CAG CCG CAT TTC AGC GAC CTG
Phe Ala Asn Thr Gly Phe Phe Arg Pro Asp Glu Ala His Gln Pro His Phe Ser Asp Leu
550      560      570      580      590      600
TTT GGG CAA ATC ATT AAC GCC GGG CAA GGG GAA GGG CGC TAT TCG GAG CTG CTG GCG ATA
Phe Gly Gln Ile Ile Asn Ala Gly Gln Gly Glu Gly Arg Tyr Ser Glu Leu Leu Ala Ile
610      620      630      640      650      660
AAT CTG CTT GAG CAA TTG TTA CTG CGG CGC ATG GAA GCG ATT AAC GAG TCG CTC CAT CCA
Asn Leu Leu Glu Gln Leu Leu Leu Arg Arg Met Glu Ala Ile Asn Gln Ser Leu His Pro
670      680      690      700      710      720
CCG ATG GAT AAT CCG GTA CGC GAG GCT TGT CAG TAC ATC AGC GAT CAC CTG GCA GAC AGC
Pro Met Asp Asn Arg Val Arg Glu Ala Cys Gln Tyr Ile Ser Asp His Leu Ala Asp Ser
730      740      750      760      770      780
AAT TTT GAT ATC GCC AGC GTC GCA CAG CAT GTT TGC TTG TCG CCG TCG CGT CTG TCA CAT
Asn Phe Asp Ile Ala Ser Val Ala Gln His Val Cys Leu Ser Pro Ser Arg Leu Ser His
790      800      810      820      830      840
CTT TTC CGC CAG CAG TTA GGG ATT AGC GTC TTA AGC TGG CGC GAG GAC CAA CGT ATC AGC
Leu Phe Arg Gln Gln Leu Gly Ile Ser Val Leu Ser Trp Arg Glu Asn Gln Arg Ile Ser
850      860      870      880      890      900
CAG GCG AAG CTG CTT TTG AGC ACC ACC CGG ATG CCT ATC GCC ACC GTC GGT CGC AAT GTT
Gln Ala Lys Leu Leu Leu Ser Thr Thr Arg Met Pro Ile Ala Thr Val Gly Arg Asn Val
910      920      930      940      950      960
GGT TTT GAT GAC CAA CTC TAT TTC TCG CGG GTA TTT AAA AAA TGC ACC GGG GCC AGC CCG
Gly Phe Asp Asp Gln Leu Tyr Phe Ser Arg Val Phe Lys Lys Cys Thr Gly Ala Ser Pro
970      980      990      1000      1010      1020
AGC GAG TTC CGT GCC GGT TGT GAA GAA AAA GTG AAT GAT GTA GCC GTC AAG TTG TCA TAA
Ser Glu Phe Arg Ala Gly Cys Glu Glu Lys Val Asn Asp Val Ala Val Lys Leu Ser END
1030      1040      1050      1060      1070      1080      1090      1100
TTGGTAACGAATCAGACAATTGACGGCTTGACGGAGTAGCATAGGTTTGAGAAATCCCTGCTTCGTCATTGACAGG
1110      1120
CACATTATGCAAGCATGGC

```

Figure 2. The nucleotide sequence of the *araC* gene and its corresponding protein product. The above figure combines sequencing data derived using the sequencing strategy shown in Figure 1 with previous sequence data (7). The nucleotide sequence is numbered in a 5' to 3' direction with the *araC* mRNA start site taken as 1. The inferred amino acid sequence is shown below the nucleotide sequence.

Figure 3. Autoradiograms of 8% polyacrylamide gels showing the position of the nonsense mutation araC5. (Left) The wild-type sequence from position 422 to 435. (Right) The araC5 sequence from position 422 to 435.

analysis of the mutation by Irr and Englesberg (19). The location of araC5, as determined by genetic mapping corresponds very closely to its location as determined by the DNA sequence analysis. Previous genetic mapping studies had placed araC5 at a point 40% of the distance through the araC gene (19). Combining the genetic mapping data with the known length of araC shown in Figure 2, we calculated that araC5 would be found near nucleotide position 400 which is very close to the actual location at position 429.

DNA sequence of a deletion mutant

Deletion 719 does not recombine with any known mutation in araC, removes the operator locus in the araBAD controlling site region and leaves the araBAD promoter intact (20). The $\Delta(\text{araC0})719$ allele was cloned by *in vivo* recombination onto the plasmid pAH15 (11). The resultant plasmid, pCH11, was used to determine the sequence changes caused by the deletion. Plasmid pCH11 was restricted with BamH I, and the resulting 5'-ends labeled with (γ - ^{32}P) ATP and polynucleotide kinase. After secondary restriction with Hinf I, the purified BamH I-Hinf I fragment was sequenced (see Figure 4). By comparison of the sequence shown in Figure 4 with the sequence shown in Figure 2 it can be concluded that $\Delta(\text{araC0})719$ begins at nucleotide position -23 and ends at 949. Thus, this deletion removes all but the terminal 22 codons of araC.

In summary, we have completed the nucleotide sequence of the araC gene.

Figure 4. Autoradiogram of an 8% polyacrylamide gel showing the sequence of ara deletion 719.

An open reading frame containing 292 codons is terminated by a single ochre codon (TAA). Proof that the open reading frame codes for araC protein was obtained by sequencing the nonsense mutant, araC5. In addition, the ara deletion 719 was found to terminate within the coding region of the gene. The determination of the DNA sequence of the araC gene will facilitate further analysis of the structure-function relationships involved in the regulatory events mediated by araC protein.

ACKNOWLEDGEMENTS

We would like to thank Laurel Heffernan for reading the manuscript and Sabi Israel for excellent technical assistance. This work was supported by National Science Foundation Grant PCM 78-26255. A.H.H. was the recipient of an NIH postdoctoral fellowship GM6914. L.G.C. was supported by USPHS-National Research Award No. 07104. C.G.M. was supported by USPHS Training Grant No. 07185. G.W. was supported by an American Cancer Society Faculty Research Award.

REFERENCES

1. Englesberg, E. and Wilcox, G. (1974) Ann. Rev. Genet. 8, 219-242

2. Lee, N. (1978) in *The Operon*, Miller, J.H. and Reznikoff, W.S. Eds. pp. 389-409, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
3. Lee, N., Wilcox, G., Gielow, W., Arnold, J., Cleary, P. and Englesberg, E. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 634-638
4. Casadaban, M. (1976) *J. Mol. Biol.* 104, 557-566
5. Wilcox, G., Boulter, J. and Lee, N. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 3635-3639
6. Smith, B. and Schleif, R. (1978) *J. Biol. Chem.* 253, 6931-6933
7. Cass, L.G., Horwitz, A.H., Miyada, C.G., Greenfield, L. and Wilcox, G. (1980) *Mol. Gen. Genet.* 180, No. 1 in press
8. Sato, S., Hutchison, C.A., III and Harris, I.I. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 542-546
9. Murray, K., Hughes, S.G., Brown, J.S. and Bruce, S. (1976) *Biochem. J.* 159, 317-322
10. Johnson, R.A. and Walseth, T.F. (1979) *Adv. Cyclic Nucleotide Res.* 10, 135-167
11. Horwitz, A.H., Heffernan, L., Cass, L., Miyada, C.G. and Wilcox, G. (1980) *Mol. Gen. Genet.* 179, No. 3 in press
12. Greenfield, L., Boone, T. and Wilcox, G. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 4724-4728
13. Norgard, M., Emigholz, K. and Monohan, J. (1979) *J. Bacteriol.* 138, 270-272
14. Maxam, A.M. and Gilbert, W. (1980) in *Methods in Enzymology*, Grossman, L. and Moldave, K. Eds., Vol. 65, pp. 499-560, Academic Press, New York.
15. Smith, H.O. and Birnstiel, M.L. (1976) *Nucl. Acids Res.* 3, 2387-2398
16. Sanger, F. and Coulson, A.R. (1978) *FEBS Lett.* 87, 107-110
17. Queen, C.L. and Korn, L.J. (1980) in *Methods in Enzymology*, Grossman, L. and Moldave, K. Eds., Vol. 65, pp. 595-609, Academic Press, New York.
18. Wilcox, G., Clemetson, K.J., Cleary, P., and Englesberg, E. (1974) *J. Mol. Biol.* 85, 589-602
19. Irr, J. and Englesberg, E. (1970) *Genetics* 65, 27-39
20. Englesberg, E., Squires, C. and Meronk, F., Jr. (1969) *Proc. Natl. Acad. Sci. U.S.A.* 62, 1100-1107