Volume 11 Number 24 1983 Nucleic Acids Research

Nucleotide sequence of the *Escherichia coli* xanthine-guanine phosphoribosyl transferase gene

Dickson Pratt and Suresh Subramani

Department of Biology, B-022, Bonner Hall, University of California, San Diego, La Jolla, CA 92093, USA

Received 6 October 1983; Revised and Accepted 7 November 1983

ABSTRACT

The Escherichia coli gene coding for the enzyme xanthine-guanine phosphoribosyl transferase (gpt) has been widely used as a dominant selectable marker in a variety of mammalian cells. We have determined the complete nucleotide sequence of the 1057 base pair (bp) segment of DNA containing this gene. The coding sequence for the enzyme is 456 nucleotides long and can code for a 152 amino acid (16.9 Kd) polypeptide. A comparison of the amino acid sequence of the bacterial enzyme with that of the mammalian hypoxanthine-guanine phosphoribosyl transferase (hprt) reveals no significant homology between the two polypeptides.

INTRODUCTION

Xanthine-guanine phosphoribosyl transferase (EC 2.4.2.22) from E. coli is a purine salvage enzyme which converts either xanthine or guanine, using phosphoribosylpyrophosphate (PRPP) as the other substrate, to XMP or GMP, respectively. Besides the gpt enzyme, E. coli also has adenine (1,2) and hypoxanthine guanine phosphoribosyltransferases (3) to salvage purines. The gpt and hprt enzymes are distinct and are coded for by different genes (4). Recently, the product of the gpt gene has been purified from E. coli harboring a plasmid that overproduces gpt and the enzyme has been reported to be a trimer with a subunit molecular weight of about 18.5 Kd(5).

The gpt gene has previously been expressed from simian virus 40-pBR322 hybrid plasmid vectors (6,7) and has been extremely useful as a dominant selectable marker for mammalian cells (6). The basis for this selection resides in the unique ability of the gpt gene product to salvage xanthine under appropriate selective conditions to produce XMP. The mammalian hprt enzyme lacks tne ability to utilize xanthine.

In the course of our studies using this gene to analyze recombination in mammalian cells, it became necessary to obtain sequence information from the 1.06 kb DNA segment within which the gene had been previously localized (6). We report here the complete nucleotide sequence of this 1057 bp DNA segment

that includes the gpt coding region as well as other upstream and downstream flanking sequences. The coding region for the enzyme is 456 nucleotides long and can code for a 16.9 Kd polypeptide subunit. The putative procaryotic transcription initiation and termination signals have also been identified.

MATERIALS AND METHODS

Enzymes

Restriction endonucleases and T4 polynucleotide kinase were purchased from New England Biolabs. Terminal transferase was from PL Biochemicals.

DNA Preparation

The 1.06 Kb DNA fragment containing the gpt gene had been cloned previously in the SV40-pBR3222 vectors pSV2gpt and pSV1GT7gpt whose structures are described in detail elsewhere (6). These DNAs were prepared by banding twice in CsCl-ethidium bromide gradients. The DNA fragments to be sequenced were obtained from one of the above two plasmids by digestion with appropriate restriction enzymes. The fragments were isolated from agarose gels using a modification of the glass powder method (8). Figure 1A shows a partial restriction map of the gene and Figure 1B shows the actual fragments used for the sequencing.

Sequence Determination

The Maxam-Gilbert procedure (9) was used for DNA sequencing. The DNA fragments labeled on the 5' end were prepared using T4 polynucleotide kinase and $\gamma[^{32}P]$ ATP (9). Alternatively, DNA fragments were labeled on the 3' end with terminal transferase and either $\alpha[^{32}P]$ labeled cordecypin triphosphate or dideoxyadenosine triphosphate according to Tu and Cohen (10). Labeled fragments were desalted on small Sephadex G-50 columns, before they were used for sequencing reactions, which were done as described by Maxam and Gilbert (9), except that for the G + A reaction, the use of 4% formic acid gave markedly better results than formic acid adjusted to pH 2.0 with piperidine.The sequencing gels were prepared as described by Maxam and Gilbert except that the gels were made in 0.75x Tris-borate-EDTA (running buffer was 1x Tris-borate-EDTA) to speed up the electrophoresis. Both DNA strands were sequenced independently between nucleotides 304-1057 as shown in Figure 1B. The region from nucleotide 120-304 was sequenced only on one strand to reconfirm the sequence published by Mulligan and Berg (7) for the first 320 nucleotides.

RESULTS AND DISCUSSION

DNA and Protein Sequence

An *E. coli* DNA fragment (1.06 Kb) containing the gpt gene had previously

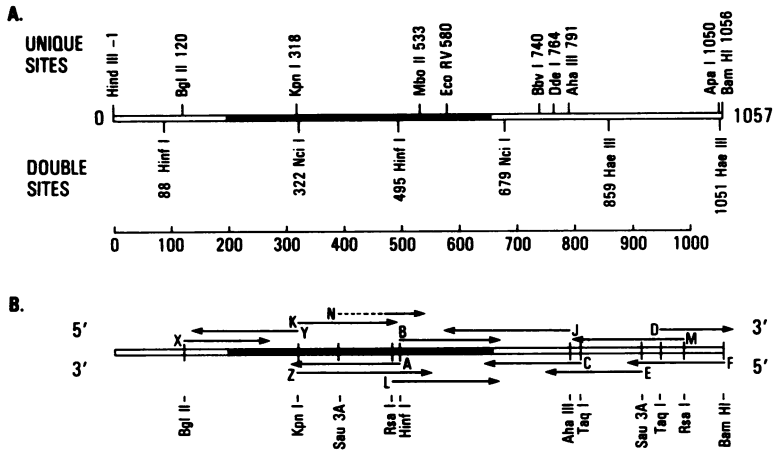

Figure 1A: Partial restriction map of the gpt DNA segment. Restriction enzymes that cut the segment once are indicated above the line. Enzymes that cut the segment twice are below the line. The shaded area represents the amino acid coding region of the gpt gene. The complete restriction map is in Table 1.
Figure 1B: Fragments used to determine the DNA sequence. Arrows show the direction and reading distance of each fragment used. Fragments shown above and below the DNA segment provided the sequence of the complementary strands. The sequence of the first 120 nucleotides was not determined because this has been published earlier (7). The dotted line extending from fragment N indicates that the sequence corresponding to the dotted portion of the fragment was not read because it had already been obtained by sequencing fragments A,Z, and K.

been cloned using HindIII and BamH1 linkers in SV40-pBR322 vectors (6). This segment of DNA shown in Figure 1A served as a convenient starting point for the sequence determination. The fragments used for the sequencing are shown in Figure 1B. The nucleotide sequence of this 1057 bp DNA segment is snown in Figure 2. It includes 198 nucleotides upstream of the initiation codon for gpt, the 456 nucleotide coding region for the enzyme and another 400 nucleotides downstream of the termination codon. The coding region begins with the ATG codon at nucleotide 199 and ends at the TAA at nucleotide 655. Our data matches exactly with the DNA sequence of the first 320 nucleotides determined earlier by Mulligan and Berg (7).

Inspection of the sequence in Figure 2 shows that there are two potential AUG codons from which translation might initiate in the message - the first corresponding to the one at nucleotide 199 and the second at nucleotide 229. Both of these codons are in the same translation reading frame. We have localized the authentic initiation codon to the one starting at nucleotide 199

```
        10         20         30         40         50         60
AGCTTGGACA CAAGACAGGC TTGCGAGATA TGTTTGAGAA TACCACTTTA TCCCGCGTCA

        70         80         90        100        110        120
GGGAGAGGCA GTGCGTAAAA AGACGCGGAC TCATGTGAAA TACTGGTTTT TAGTGCGCCA

       130        140        150        160        170        180
GATCTCTATA ATCTCGCGCA ACCTATTTTC CCCTCGAACA CTTTTTAAGC CGTAGATAAA

       190        200        210        220        230        240
CAGGCTGGGA CACTTCACAT GAGCGAAAAA TACATCGTCA CCTGGGACAT GTTGCAGATC

       250        260        270        280        290        300
CATGCACGTA AACTCGCAAG CCGACTGATG CCTTCTGAAC AATGGAAAGG CATTATTGCC

       310        320        330        340        350        360
GTAAGCCGTG GCGGTCTGGT ACCGGGTGCG TTACTGGCGC GTGAACTGGG TATTCGTCAT

       370        380        390        400        410        420
GTCGATACCG TTTGTATTTC CAGCTACGAT CACGACAACC AGCGCGAGCT TAAAGTGCTG

       430        440        450        460        470        480
AAACGCGCAG AAGGCGATGG CGAAGGCTTC ATCGTTATTG ATGACCTGGT GGATACCGGT

       490        500        510        520        530        540
GGTACTGCGG TTGCGATTCG TGAAATGTAT CCAAAAGCGC ACTTTGTCAC CATCTTCGCA

       550        560        570        580        590        600
AAACCGGCTG GTCGTCCGCT GGTTGATGAC TATGTTGTTG ATATCCCGCA AGATACCTGG

       610        620        630        640        650        660
ATTGAACAGC CGTGGGATAT GGGCGTCGTA TTCGTCCCGC CAATCTCCGG TCGCTAATCT

       670        680        690        700        710        720
TTTCAACGCC TGGCACTGCC GGGCGTTGTT CTTTTTAACT TCAGGCGGGT TACAATAGTT

       730        740        750        760        770        780
TCCAGTAAGT ATTCTGGAGG CTGCATCCAT GACACAGGCA AACCTGAGCG AAACCCTGTT

       790        800        810        820        830        840
CAAACCCCGC TTTAAAACATC CTGAAACCTC GACGCTAGTC CGCCGCTTTA ATCACGGCGC

       850        860        870        880        890        900
ACAACCGCCT GTGCAGTCGG CCCTTGATGG TAAAACCATC CCTCACTGGT ATCGCATGAT

       910        920        930        940        950        960
TAACCGTCTG ATGTGGATCT GGCGCGGCAT TGACCCACGC GAAATCCTCG ACGTCCAGGC

       970        980        990       1000       1010       1020
ACGTATTGTG ATGAGCGATG CCGAACGTAC CGACGATGAT TTATACGATA CGGTGATTGG

      1030       1040       1050
CTACCGTGGC GGCAACTGGA TTTATGAGTG GGCCCCG
```

Figure 2. Nucleotide sequence of the gpt DNA segment. The sequence of the 1057 bp DNA segment carrying the gpt gene is shown. The initiation and termination codons marking the boundaries of the gpt coding sequence are underlined. The sequence of the first 120 bp is from Mulligan and Berg (7).

Table 1. Restriction Sites in the gpt Segment

Name	Sequence	No. of Sites	Location
AhaIII	TTTAAA	1	791
ApaI	GGGCCC	1	1050
BbvI	GCTGC	1	740
BglII	AGATCT	1	120
DdeI	CTGAG	1	764
EcoRV	GATATC	1	580
KpnI	GGTACC	1	318
MboII	TCTTC	1	533
HaeIII	GGCC	2	859,1051
HinfI	GACTC,GATTC	2	88,495
NciI	CCGGG	2	322,679
AluI	AGCT	3	1,382,407
FokI	CATCC	3	744,797,877
HphI	TCACC,GGTGA	3	218,527,1012
RsaI	GTAC	3	319,482,987
Sau96I	GGCCC,GGGCC	3	859,1050,1051
SfaNI	GATGC,GCATC	3	267,743,977
XhoII	AGATCT,AGATCC,GGATCT	3	120,236,915
Fnu4HI	GCTGC,GCCGC,GCGGC	4	740,822,924,1029
HgaI	GCGTC,GACGC	4	55,82,623,811
Sau3A	GATC	4	121,237,388,916
TaqI	TCGA	4	154,362,809,948
BstNI	CCTGG,CCAGG	5	221,465,596,669,955
HpaII	CCGG	5	322,476,544,647,679
MnlI	GAGG,CCTC	6	65,152,737,807,881,946
ScrFI	CCTGG,CCGGG,CCAGG	7	221,322,465,596,669,679,955
HhaI	GCGC	8	115,136,337,402,425,517,837,922
ThaI	CGCG	8	54,84,135,338,403,424,923,938

Restriction Sites Not Found

AccI,AvaI,AvaII,BalI,BclI,BglI,BssHII,BstEII,ClaI,EcoRI,HaeII,HgiAI, HincII,HpaI,MluI,MstI,MstII,NaeI,NarI,NcoI,NdeI,NruI,PstI,PvuI,PvuII, SalI,SmaI,SphI,SstI,SstII,StuI,TthlllI,XbaI,XhoI,XmaIII,XmnI.

because the DNA segment from nucleotide 197-1057 produced authentic gpt in mammalian cells when expressed as part of a eucaryotic transcription unit, but the DNA segment from nucleotide 214-1057 did not express gpt (D. Peabody; personal communication). The 152 amino acid sequence of the gpt gene product is shown in Figure 3. The predicted size of the polypeptide subunit (16.9 Kd) agrees reasonably well with the 18.5 Kd value reported earlier (5). There is a second open reading frame starting at nucleotide 749 and continuing past the end of the DNA sequence shown, but it is not known whether this codes for any protein *in vivo*. This second reading frame could not code for gpt for a variety of reasons. Mulligan and Berg (7) have demonstrated that the deletion of the first 318 nucleotides in Figure 2 (up to the KpnI site) resulted in the complete loss of gpt activity in bacteria and in mammalian cells where the gpt

1 20

Met Ser Glu Lys Tyr Ile Val Thr Trp Asp Met Leu Gln Ile His Ala Arg Lys Leu Ala
Ser Arg Leu Met Pro Ser Glu Gln Trp Lys Gly Ile Ile Ala Val Ser Arg Gly Gly Leu
Val Pro Gly Ala Leu Leu Ala Arg Glu Leu Gly Ile Arg His Val Asp Thr Val Cys Ile
Ser Ser Tyr Asp His Asp Asn Gln Arg Glu Leu Lys Val Leu Lys Arg Ala Glu Gly Asp
Gly Glu Gly Phe Ile Val Ile Asp Asp Leu Val Asp Thr Gly Gly Thr Ala Val Ala Ile
Arg Glu Met Tyr Pro Lys Ala His Phe Val Thr Ile Phe Ala Lys Pro Ala Gly Arg Pro
Leu Val Asp Asp Tyr Val Val Asp Ile Pro Gln Asp Thr Trp Ile Glu Gln Pro Trp Asp
Met Gly Val Val Phe Val Pro Pro Ile Ser Gly Arg

Figure 3. Amino acid sequence of the gpt gene product.

gene segment was linked to a functional SV40 promoter and transcription unit. This localizes a part of the gpt coding sequence within the first 318 nucleotides and not within the second open reading frame. Moreover, it is known that the gpt protein coded for by the DNA segment in Figure 2 is exactly the same size as authentic E. coli gpt when expressed from different expression vectors in bacterial or mammalian cells (5,7). Thus, the entire gpt gene must be present within the sequence shown in Figure 2. The second open reading frame is not only too small to code for authentic gpt, but it is also an incomplete reading frame that continues past the end of the DNA segment in Figure 2, and it is therefore quite unlikely that it could code for authentic size gpt when expressed from different procaryotic and eucaryotic expression vectors. A comparison of the amino acid sequences of gpt and human hprt (13) revealed no significant homology between the two polypeptides.

Regulatory Signals

Examination of the sequence upstream of the coding sequence reveals a potential Pribnow box (11) (nucleotides 127-132) and other promoter-associated sequences about 20-25 nucleotides upstream of the Pribnow box (7). The transcription initiation site for the gpt mRNA is unknown. However, if transcription starts downstream of the Pribnow box as predicted, then the correct initiation codon would be the first one in the message. Just downstream from the TAA termination codon is a sequence AACGCCTGGCACTGCCGGGCGTT (nucleotides 665-687) which is potentially capable of forming a stem-loop structure. This feature, along with the sequence of five T residues (nucleotides 692-696) following it, resembles the transcription termination signal found in procaryotes (12).

ACKNOWLEDGEMENTS

While this manuscript was being prepared, we became aware that K. Richardson, J. Fostel and T. Skopek had also determined the sequence of the gpt gene independently. We thank Tom Skopek for exchanging his sequence information with ours. We also thank R. Doolittle and Dan Donoghue for advice and for help with the computer. This work was supported by grants from NIH and tne Searle Scholars Program.

REFERENCES

1. Hochstadt-Ozer, J. and Stadtman, E.R. (1971) J. Biol. Chem. 246, 5294-5303.
2. Hochstadt-Ozer, J. and Stadtman, E.R. (1971) J. Biol. Chem. 246, 5312-5320.
3. Martin, W.R. and Yang, R.R. (1972) Biochem. Biophys. Res. Commun. 48, 1641-1648.
4. Holden, J.A., Harriman, P.D. and Wall, J.D. (1976) J. Bacteriol. 126, 1141-1148.
5. Liu, S.W. and Milman, G. (1983) J. Biol. Chem. 258, 7469-7475.
6. Mulligan, R.C. and Berg, P. (1980) Science, 209, 1422-1427.
7. Mulligan, R.C. and Berg, P. (1981) Mol. Cell. Biol. 1, 449-459.
8. Vogelstein, B. and Gillespie, D. (1979) Proc. Natl. Acad. Sci. USA 76, 615-619.
9. Maxam, A. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
10. Tu, C.D. and Cohen, S.N. (1980) Gene 10, 177-183.
11. Pribnow, D. (1975) Proc. Natl. Acad. Sci. USA 72, 784-788.
12. Tinoco, Jr., I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) Nature New Biol. 246, 40-41.
13. Jolly, D.J., Okayama, H., Berg, P., Esty, A.C., Filpula, D., Bohlen, P., Johnson, G.G., Shively, J.E., Hunkapillar, T. and Friedmann, T. (1983) Proc. Natl. Acad. Sci. USA 80, 447-481.