Complete nucleotide sequence of the *E. coli* *glyA* gene

Michael D.Plamann, Lorraine T.Stauffer, Mark L.Urbanowski and George V.Stauffer

Department of Microbiology, University of Iowa, Iowa City, IA 52242, USA

Received 29 December 1982; Revised and Accepted 28 February 1983

ABSTRACT

The nucleotide sequence of the *Escherichia coli* *glyA* gene has been determined. The amino acid sequence predicted from the DNA sequence consists of 417 residues. After the coding region there is a 185 nucleotide sequence preceding the proposed transcription termination region for the *glyA* gene. This region is preceded by a G-C rich sequence that could form a stable stem-loop structure once transcribed, followed by an A-T rich sequence within which transcription appears to terminate. There is a long region of dyad symmetry and numerous smaller symmetrical regions between the site of translation termination and the proposed transcription termination region. These stem-loop structures show remarkable homology with intercistronic elements of other prokaryotic operons and may play a role in the regulation of *glyA* gene expression.

INTRODUCTION

Serine hydroxymethyltransferase (SHMT), the *glyA* gene product, is responsible for the conversion of serine to glycine and 5,10-methylene-tetrahydrofolate. This reaction is a major source of one-carbon units, and also fulfills the cells need for glycine (1).

SHMT regulation is complex. Serine, glycine, methionine, thymine, purines and folates all appear to be involved in regulation of *glyA* gene expression, but the mechanism(s) of their involvement is unknown (2-6).

As a first step toward understanding the molecular mechanism of *glyA* gene regulation we cloned the *E. coli* *glyA* gene onto multicopy plasmid vectors (7). A preliminary report described the DNA sequence and a biochemical analysis of the *glyA* control region (8). In this paper we present the complete DNA sequence of the *glyA* gene and its 3' flanking region.

MATERIALS AND METHODS

Bacteria and Plasmids. The *E. coli* K12 strain GS245 (*pheA905 thi ΔglyA araD139 ΔlacU169 strA*) was used in all transformations

and preparations of plasmid DNA (7). Plasmids pGS1, pGS27 and pGS29 contain the *E. coli glyA* gene and have been described (7).

Plasmid isolation. Plasmid DNA was prepared by described methods (9-12).

DNA sequence analysis. The DNA sequencing procedure of Maxam and Gilbert (13) was used with the modifications of Smith and Calvo (14). Gel electrophoresis was carried out according to Sanger and Coulson (15).

3' S1 mapping. The S1 mapping procedure of Weaver and Weissman (16) was used, with slight modification (8), to identify the possible transcription termination region. A 248 base pairs (bp) *TaqI* DNA fragment that spans the *glyA* transcription termination region was labelled at the 3' end using the large fragment of *E. coli* DNA polymerase I and [α -³²P]dCTP (17). The ³²P-labelled coding strand (about 1 μ g) was purified electrophoretically (13), hybridized to total cellular RNA (about 10 μ g) and digested with varying amounts of S1 nuclease (10-100 U) for 40 minutes at 30 C. The products of this digestion were prepared for electrophoresis and run adjacent to a DNA sequencing ladder of the same fragment.

Enzymes and Chemicals. Restriction endonucleases, large fragment of DNA polymerase I, bacterial alkaline phosphatase, and S1 nuclease were purchased from Bethesda Research Laboratory (Gaithersburg, MD) or New England Biolabs (Beverly, MA). T4 polynucleotide kinase was obtained from P-L Biochemicals (Milwaukee, WI). All restriction endonuclease digestion conditions were as described by the manufacturers. [γ -³²P]ATP and [α -³²P]dCTP were from Amersham (Arlington Heights, IL). All other chemicals were reagent grade and commercially available.

RESULTS

The *glyA* gene was initially isolated from *E. coli* K12 on a 13 kilo-base pairs (kb) *EcoRI* fragment (7). Subsequent subcloning experiments localized the *glyA* gene to a 1.9 kb *HpaII-PvuII* DNA fragment (8). A physical map of this fragment is presented in Fig. 1 along with the DNA sequencing strategy. The DNA sequence of both strands was determined and all restriction sites were overlapped. Fig. 2 shows the nucleotide sequence of the *E. coli* K12 *glyA* gene along with the deduced amino acid sequence.

Amino acid sequence and composition. The DNA sequence presented in Fig. 2 has a major open reading frame extending from position 68 to 1318. This open reading frame codes for a 417 amino acids long polypeptide,

Figure 1. Nucleotide sequence determinations and restriction endonuclease recognition sites used to establish the *glyA* sequence. Arrows indicate the extent of each sequence determination.

Table 1. Amino Acid Composition of SHMT

Ala	47	Gln	14	Leu	31	Ser	17
Arg	14	Glu	29	Lys	29	Thr	18
Asn	18	Gly	41	Met	12	Trp	3
Asp	21	His	13	Phe	11	Tyr	20
Cys	3	Ile	21	Pro	19	Val	36

Total number of residues = 417; M_r = 45,265

Table 2. Codon Usage in *glyA*

Phe	TTT	3	Ser	TCT	4	Tyr	TAT	5	Cys	TGT	1
Phe	TTC	8	Ser	TCC	7	Tyr	TAC	15	Cys	TGC	2
Leu	TTA	1	Ser	TCA	0	End	TAA	1	End	TGA	0
Leu	TTG	1	Ser	TCG	0	End	TAG	0	Trp	TGG	3
Leu	CTT	0	Pro	CCT	4	His	CAT	3	Arg	CGT	8
Leu	CTC	2	Pro	CCC	0	His	CAC	10	Arg	CGC	6
Leu	CTA	0	Pro	CCA	1	Gln	CAA	2	Arg	CGA	0
Leu	CTG	27	Pro	CCG	14	Gln	CAG	12	Arg	CGG	0
Ile	ATT	4	Thr	ACT	6	Asn	AAT	1	Ser	AGT	0
Ile	ATC	17	Thr	ACC	12	Asn	AAC	17	Ser	AGC	6
Ile	ATA	0	Thr	ACA	0	Lys	AAA	26	Arg	AGA	0
Met	ATG	12	Thr	ACG	0	Lys	AAG	3	Arg	AGG	0
Val	GTT	19	Ala	GCT	17	Asp	GAT	13	Gly	GGT	24
Val	GTC	5	Ala	GCC	10	Asp	GAC	8	Gly	GGC	17
Val	GTA	6	Ala	GCA	4	Glu	GAA	21	Gly	GGA	0
Val	GTG	6	Ala	GCG	16	Glu	GAG	8	Gly	GGG	0

TTTCCGGTTGCAAGCTCTTTATTCTCCAAAGCCTTGCCTAGCCTGAAGGTAATCGTTTGCGTAAATTCCTTTGTCAAGAC

CTGTTATCGCACAAATGATTTCGGTTTATACTGTTGCGCGTTGTCCAACAGGACCGCCTATAAAGGCCAAAAATTTATTGTT
-35 Pribnow box
region sequence

5
Met Leu Lys Arg Glu Met Asn Ile Ala Asp Tyr Asp Ala Glu Leu
AGCTGAGTCAGGAGATGCGG ATG TTA AAG CGT GAA ATG AAC ATT GCC GAT TAT GAT GCC GAA CTG
Shine-Dalgarno
sequence

25
Trp Gln Ala Met Glu Gln Glu Lys Val Arg Gln Glu Glu His Ile Glu Leu Ile Ala Ser
TGG CAG GCT ATG GAG CAG GAA AAA GTA CGT CAG GAA GAG CAC ATC GAA CTG ATC GCC TCC
150

45
Glu Asn Tyr Thr Ser Pro Arg Val Met Gln Ala Gln Gly Ser Gln Leu Thr Asn Lys Tyr
GAA AAC TAC ACC AGC CCG CGC GTA ATG CAG GCG CAG GGT TCT CAG CTG ACC AAC AAA TAT

65
Ala Glu Gly Tyr Pro Gly Lys Arg Tyr Tyr Gly Gly Cys Glu Tyr Val Asp Ile Val Glu
GCT GAA GGT TAT CCG GGC AAA CGC TAC TAC GGC GGT TGC GAG TAT GTT GAT ATC GTT GAA

85
Gln Leu Ala Ile Asp Arg Ala Lys Glu Leu Phe Gly Ala Asp Tyr Ala Asn Val Gln Pro
CAA CTG GCG ATC GAT CGT GCG AAA GAA CTG TTC GGC GCT GAC TAC GCT AAC GTC CAG CCG
300

105
His Ser Gly Ser Gln Ala Asn Phe Ala Val Tyr Thr Ala Leu Leu Glu Pro Gly Asp Thr
CAC TCC GGC TCC CAG GCT AAC TTT GCG GTC TAC ACC GCG CTG CTG GAA CCA GGT GAT ACC

125
Val Leu Gly Met Asn Leu Ala His Gly Gly His Leu Thr His Gly Ser Pro Val Asn Phe
GTT CTG GGT ATG AAC CTG GCG CAT GGC GGT CAC CTG ACT CAC GGT TCT CCG GTT AAC TTC
450

145
Ser Gly Lys Leu Tyr Asn Ile Val Pro Tyr Gly Ile Asp Ala Thr Gly His Ile Asp Tyr
TCC GGT AAA CTG TAC AAC ATC GTT CCT TAC GGT ATC GAT GCT ACC GGT CAT ATC GAC TAC

165
Ala Asp Leu Glu Lys Gln Ala Lys Glu His Lys Pro Lys Met Ile Ile Gly Gly Phe Ser
GCC GAT CTG GAA AAA CAA GCC AAA GAA CAC AAG CCG AAA ATG ATT ATC GGT GGT TTC TCT

185
Ala Tyr Ser Gly Val Val Asp Trp Ala Lys Met Arg Glu Ile Ala Asp Ser Ile Gly Ala
GCA TAT TCC GGC GTG GTG GAC TGG GCG AAA ATG CGT GAA ATC GCT GAC AGC ATC GGT GCT
600

205
Tyr Leu Phe Val Asp Met Ala His Val Ala Gly Leu Val Ala Ala Gly Val Tyr Pro Asn
TAC CTG TTC GTT GAT ATG GCG CAC GTT GCG GGC CTG GTT GCT GCT GGC GTC TAC CCG AAC

225
Pro Val Pro His Ala His Val Val Thr Thr Thr His Lys Thr Leu Ala Gly Pro Arg
CCG GTT CCT CAT GCT CAC GTT GTT ACT ACC ACC ACT CAC AAA ACC CTG GCG GGT CCG CGC
750

245
Gly Gly Leu Ile Leu Ala Lys Gly Gly Ser Glu Glu Leu Tyr Lys Lys Leu Asn Ser Ala
GGC GGC CTG ATC CTG GCG AAA GGT GGT AGC GAA GAG CTG TAC AAA AAA CTG AAC TCT GCC

Figure 2. The nucleotide and deduced amino acid sequence of the *E. coli* *glyA* gene. The two in phase translation termination codons are indicated by bars over the sequence.

the presumed *glyA* gene product. At the N-terminus of this polypeptide the predicted AUG initiator codon is preceded by a good Shine-Dalgarno sequence (18) and at the C-terminus one finds two in phase termination codons at positions 1319 and 1331 (Figure 2).

The amino acid composition of the *glyA* gene product, as deduced from the DNA sequence, is presented in Table 1. The predicted molecular

Figure 3. Location of the 3' end of *glyA* mRNA. The coding strand of the 248 bp *Taq*I fragment containing the 3' flanking region of the *glyA* gene was labelled with ^{32}P at the 3'-terminus, hybridized to total cellular RNA, treated with S1 nuclease, denatured, and the S1 nuclease resistant DNA was electrophoresed adjacent to a sequencing ladder of the same DNA fragment. The length of the protected DNA fragment corresponds to the distance from the nucleotide encoding the 3' end of *glyA* mRNA to the 3' end of the labelled fragment. S1 nuclease treatment was with 100 U, lane a; 50 U, lane b; or 10 U, lane c.

weight (M_r) of 45,265 is in good agreement with the 46,500 value estimated from SDS-polyacrylamide gel electrophoresis (8). The codon usage for the *glyA* gene is presented in Table 2.

Location of the 3' end of *glyA* mRNA. We used the S1 mapping

```

      T C
      T  G
      C - G
      C - G
      T - A
      T - A
      C - G
      G - C
      G - C
      A - T
ATCA - TTCTACGTT

```

Figure 4. The proposed *glyA* termination region. The bar indicates the major 3' termini determined from Figure 3.

procedure of Weaver and Weissman to determine the location of the 3' end of *glyA* mRNA (16 and Materials and Methods). The results of the S1 mapping gel are shown in Figure 3. The transcription termination region deduced from these results is shown in Figure 4.

Dyad symmetries and transcription termination. The DNA sequence shown in Fig. 2 was analyzed for regions of dyad symmetry using the computer program of Queen and Korn (19). Although no major symmetries were found at the 5' end of the *glyA* gene or within the coding region, considerable regions of dyad symmetry occur between the translation termination site and the proposed transcription termination region (Figure 5).

DISCUSSION

The DNA sequence of the *glyA* gene has made it possible to deduce the amino acid sequence of the protein (Figure 2). An analysis of the

Figure 5. Comparison of the nucleotide sequence distal to the *E. coli* *glyA*, *trpR* and *glnS* genes. The sequences are aligned for homology. The *glyA* sequence is used as the standard. A dash in the sequence indicates the absence of a base. Regions of nonhomology are indicated by small letters. The orientation of the homologous sequence at the end of *glnS* is opposite that of *glyA* and *trpR*. Arrows above the sequence indicate regions of dyad symmetry. The nucleotide sequence for *trpR* is from references 32 and 33. The nucleotide sequence for *glnS* is from reference 34.

codon usage is presented in Table 2. The very non-random pattern of codon usage (i.e., for leucine 27 of 31 residues are coded for by CTG) in glyA is similar to reported patterns in other E. coli genes (20, 21). This non-random pattern shows a strong positive correlation between major tRNA iso-accepting species and choice of codons (20). A comparison of the codon usage in the glyA gene to the codon usage in strongly and moderately to weakly expressed genes in E. coli indicates that glyA is an efficiently expressed gene (21).

In the predicted amino acid sequence of SHMT, we found the sequence Val-Val-Thr-Thr-Thr-Thr-His-Lys-Thr-Leu between amino acids 222 and 231. In rabbit liver SHMT the amino acid sequence at the pyridoxal-5'-phosphate binding site was determined to be Val-Val-Thr-Thr-Thr-His-Lys-Thr-Leu (22). The nonapeptide isolated from rabbit liver SHMT shows a remarkable resemblance to this portion of the E. coli sequence. Whether this sequence is the pyridoxal-5'-phosphate binding site in the E. coli enzyme remains to be established. We also found the sequence Lys-Pro-Lys-Met-Ile-Ile-Gly-Gly-Phe-Ser-Ala-Tyr between amino acids 166 and 177. Schirch et al. (23) determined the amino acid sequence of a cysteine containing peptide from the active site of rabbit liver SHMT. Their sequence, His-Pro-Lys-Leu-Ile-Ile-Ala-Gly-Thr-Ser-Cys-Tyr, shows considerable similarity with the above amino acid sequence from E. coli. Again, it remains to be shown that this region is part of the E. coli SHMT active site.

The translation of SHMT terminates with a GCA (Ala) codon followed by a UAA stop codon (Figure 2). About 160 bp downstream from the termination codon there is a G-C rich inverted repeat sequence followed by an A-T rich sequence (Figure 4). This structure is similar to other procaryotic transcription termination signals (24) and shows a strong resemblance to the λ phage t_{L3} terminator (25). This structure could function as part of a transcription termination signal or possibly to prevent degradation of glyA mRNA. From S1 mapping experiments transcription termination is proposed to occur within the A-T rich sequence (Figures 3 and 4).

As shown in Figure 3, several protected DNA bands were observed to occur within the A-T rich sequence. Some of these bands could be due to nibbling by the S1 nuclease at the ends of the DNA-RNA hybrid (8), or perhaps there are multiple transcription stop sites. Alternatively, since the RNA used in the hybridizations was isolated from cells it could have been nibbled by ribonucleases to generate multiple ends or processed to generate the

observed 3' termini. It is not clear which of the alternatives is correct and additional studies are necessary to resolve this question.

The DNA following the translation stop codon contains a complex set of dyad symmetries (Figure 5). The longest symmetrical region transcribed can form a very stable stem-loop structure ($\Delta G = -62.4$ kcal/mole) (26). Additional smaller dyad-symmetries occur within and around this one (Figure 5). Thus, a large number of possible secondary structures, some of which are mutually exclusive, can be formed. It is not known whether these structures are functionally significant. It is possible they prevent rapid degradation of glyA mRNA. Alternatively, RNaseIII is known to cleave large stem-loop structures in mRNA and in so doing to regulate gene expression (27-30). The 3' S1 nuclease mapping experiments, however, did not detect additional 3' terminated RNA transcripts in this region. This result suggests that the potential stem-loop structures are not RNaseIII processing sites. It is possible, however, that such processed RNA transcripts would be rapidly degraded and not detected.

Recently a novel intercistronic element was observed in three independent prokaryotic operons (31). These include the histidine transport and the histidine biosynthetic operons of S. typhimurium and the malK-lamB operon of E. coli. In each case the element consists of a long region of dyad symmetry and several smaller symmetries, some of which overlap the main one. All three stem-loop structures show a high degree of homology (about 90%), suggesting a common origin. It has been postulated that these structures function to decrease expression of distal genes in an operon. The long region of dyad symmetry following the translation stop signal for the glyA gene also shows about 90% homology with these three presumed regulatory elements. It remains to be determined whether there is a cotranscribed gene distal to the glyA gene.

A similar structure is found immediately following the fol, trpR and glnS genes of E. coli (14, 32-34). The stem-loop structures of glyA, trpR and glnS show even greater homology than the above elements (95%) (Figure 5). It is interesting that the orientation of the common genetic element is reversed at the end of the glnS gene relative to the glyA and trpR sequences (Figure 5). If these elements have a common origin, such as from an insertion-like sequence (31), it is not surprising to find both possible orientations. Their role, if any, in regulation of gene expression remains unclear. If they play a role in regulating the differential expression of genes in multicistronic operons, it may be possible the same mechanism,

with modification, can function as part of the termination signal at the end of other transcriptional units.

ACKNOWLEDGEMENTS

We thank Marcia Reeve for typing the manuscript. This investigation was supported by Public Health Service grant GM26878 from the National Institute for General Medical Sciences.

REFERENCES

1. Mudd, S. H. and Cantoni, G. L. (1964) In: Comprehensive Biochemistry, Vol. 15, Florkin, M. and Stotz, E. H. (eds). pp. 1-47.
2. Taylor, R. T., Dickerman, H. and Weissbach, H. (1966) Arch. Biochem. Biophys. 117, 405-412.
3. Mansouri, A., Decter, J. B. and Silber, R. (1972) J. Biol. Chem. 247, 348-352.
4. Stauffer, G. V., Baker, C. A. and Brenchley, J. E. (1974) J. Bacteriol. 120, 1017-1025.
5. Miller, B. A. and Newman, E. B. (1974) Can. J. Microbiol. 20, 41-47.
6. Greene, R. C. and Radovich, C. (1975) J. Bacteriol. 124, 269-278.
7. Stauffer, G. V., Plamann, M. D. and Stauffer, L. T. (1981) Gene 14, 63-72.
8. Plamann, M. D. and Stauffer, G. V. (1983) Gene (in press).
9. Clewell, D. B. (1972) J. Bacteriol. 110, 667-676.
10. Guerry, P., LeBlanc, D. J. and Falkow, S. (1973) J. Bacteriol. 116: 1064-1066.
11. Selker, E., Brown, K. and Yanofsky, C. (1977) J. Bacteriol. 129, 388-394.
12. Holmes, D. S. and Quigley, M. (1981) Anal. Biochem. 114, 193-197.
13. Maxam, A. M. and Gilbert, W. (1980) Methods in Enzymol. 65, 499-560.
14. Smith, D. R. and Calvo, J. M. (1980) Nucleic Acids Res. 8, 2255-2274.
15. Sanger, F. and Coulson, A. R. (1978) FEBS Letters 87, 107-110.
16. Weaver, R. F. and Weissman, C. (1979) Nucleic Acids Res. 7, 1175-1193.
17. Wu, R. (1970) J. Mol. Biol. 51, 501-521.
18. Shine, J. and Dalgarno, L. (1975) Nature 254, 34-38.
19. Queen, C. L. and Korn, L. J. (1980) Methods in Enzymology 65, 595-609.
20. Ikemura, T. (1981) J. Mol. Biol. 146, 1-21.
21. Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.
22. Bossa, F., Barra, D., Martini, F., Schirch, L., and Fasella, P. (1976) Eur. J. Biochem. 70, 397-401.
23. Schirch, L., Slagel, S., Barra, D., Martini, F. and Bossa, F. (1980) J. Biol. Chem. 255, 2986-2989.
24. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353.
25. Luk, K.-C. and Szybalski, W. (1982) Gene 17, 247-258.
26. Tinoco, I., Jr., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., and Gralla, J. (1973) Nature New Biol. (London) 246, 40-41.
27. Gegenheimer, P., Watson, N. and Apirion, D. (1977) J. Biol. Chem. 252, 3064-3073.

28. Robertson, H. D., Dickson, E. and Dunn, J. J. (1977) Proc. Natl. Acad. Sci. 74, 822-826.
29. Bram, R. J., Young, R. A. and Steitz, J. A. (1980) Cell 19, 393-401.
30. Barry, G., Squires, C. and Squires, C. L. (1980) Proc. Natl. Acad. Sci. 77, 3331-3335.
31. Higgins, C. F., Ames, G. F.-L., Barnes, W. M., Clement, J. M. and Hofnung, M. (1982) Nature 298, 760-762.
32. Gunsalus, R. P. and Yanofsky, C. (1980) Proc. Natl. Acad. Sci. 77, 7117-7121.
33. Singleton, C. K., Roeder, W. D., Bogosian, G., Somerville, R. L. and Weith, H. L. (1980) Nucleic Acids Res. 8, 1551-1560.
34. Yamao, F., Inokuchi, H., Cheung, A., Ozeki, H. and Söll, D. (1982) J. Biol. Chem. 257, 11639-11643.