Volume 12 Number 24 1984 Nucleic Acids Research

Complete nucleotide sequence of the *E. coli* glutathione synthetase gsh-II

Hiroshi Gushima[2], Shuhay Yasuda[3], Eiichi Soeda[3], Masami Yokota[2], Masatoshi Kondo[2] and Akira Kimura[1]

[1]Research Institute for Food Science, Kyoto University, Uji, Kyoto 611, [2]Central Research Laboratory of Yamanouchi Pharmaceutical Company, 1-1-8 Azusawa, Itabashi-ku, Tokyo 174, and [3]National Institute of Genetics, 1,111 Yata, Mishima 411, Japan

Received 24 October 1984; Accepted 7 November 1984

ABSTRACT

The nucleotide sequence of the cloned DNA, 1,478 bp in length coding for glutathione synthetase (GSH-II) of E.coli B has been determined. Amino acid and nucleotide sequence analyses have assigned the open reading frame for GSH-II, starting with the ATG near its 5' terminus. The molecular weight caluculated from the predicted amino acid sequence is 35,559 daltons, being in good agreement with that of a GSH-II subunit estimated by the SDS-PAGE method. Several signal sequences conserved in the promoter regions of E.coli were found in the non-coding regions of the gsh-II gene. They include the Shine-Dalgarno sequence, the Pribnow box and the sequence conserved in the "-35 region" with a preferable spacing from each other for an efficient transcription. Downstream from the termination codon, the inverted repeat sequences were present, followed by 6 successive T's. These structural features found in the non-coding regions have suggested to be involved in regulatory functions for the gsh-II gene expression.

INTRODUCTION

Glutathione synthetase (EC 6.3.2.3)(GSH-II), the second enzyme in the glutathione biosynthetic pathway in E.coli B, catalyzes the condensation of γ-glutamylcysteine and glycine in the presence of ATP to give glutathione. As shown previously (1-4), we have been studying the production of glutathione by E.coli cells in a bioreactor system. In order to attain a higher production of glutathione, the gene for GSH-II has been cloned into pBR322 and designated pGS200 (2). Recently the GSH-II has been purified from the cell extract of E.coli B transformed with pGS200 and well characterized enzymologically (5). An apparent molecular weight of native GSH-II was estimated to be about 152 kd by gel filtration method. This enzyme is active in a tetrameric form of the identical subunit with a molecular weight of approximately 38 kd. In the present work, we have sequenced the cloned DNA fragment, which includes the entire gsh-II gene as well as the flanking regulatory regions.

MATERIALS AND METHODS

Bacterial Strains

Strain *E.coli* B (C1001) and *E.coli* K-12 [C600(F⁻ *hsdR hsdM recA*⁺ *thr leu thi lacY supE tonA*)] were used. The strain C1001 was deficient in GSH-II activity.

Enzymes and Biochemicals

Restriction endonucleases were obtained from Takara Shuzo Co. Ltd, and Nippon Gene Co. Ltd. Nuclease P1 was from Yamasa Shoyu Co. Ltd. [α-^{32}P]dCTP was purchased from RCC Amersham. M13 cloning and sequencing kits were obtained from either RCC Amersham or Takara Shuzo Co. Ltd. Tetramethylthiuram disulfide and other reagents used for enzyme assay were purchased from Sigma Chemical Co. Ltd.

DNA sequencing procedure

pGS200 was digested with the combination of *BamHI* and *HindIII* restriction endonucleases and the resulting 1400 bp fragment was separeted from the vector by electrophoresis in a 10% low-melting-temperature agarose gel. The fragment was redigested with either *RsaI* or *HaeIII*, inserted into the corresponding the cloning sites of M13 mp8 and M13 mp9 phage vectors and subcloned according to the supplier's specifications (RCC Amercham). Alternatively, the fragment was sonicated into pieces and inserted at the *SmaI* site of M13 mp8 DNA after processed with nuclease P1 and T4 DNA polymerase to generate flush ends. A shotgun library was constructed by the modified method (Yasuda et al., submitted) of Deininger (8). The resulting recombinant phage DNA was sequenced by the "dideoxy sequencing method" of Sanger et al (9).

DNA sequence analyses were performed with DNASIS (Hitachi SK Co.)

Purification of GSH-II and amino acid sequence determination

GSH-II was purified from *E.coli* C600 cells transformed with pGS200. The purification method was essentially as reported previously (5) except that FPLC system with Mono Q column (Pharmacia) was used in the final step. The amino-terminal sequence of the purified GSH-II at one a nmol scale was determined by the use of Model 470A protein sequencer (Applied Biosystems).

RESULTS

Subcloning of the gsh-II gene

pGS200 was 6900 bp in length and contained the gsh-II gene derived from *E.coli* B chromosomal DNA fragment (2600 bp) inserted at the *HindIII* site of pBR322 (2). To obtain a plasmid containg the shorter gsh-II gene, we

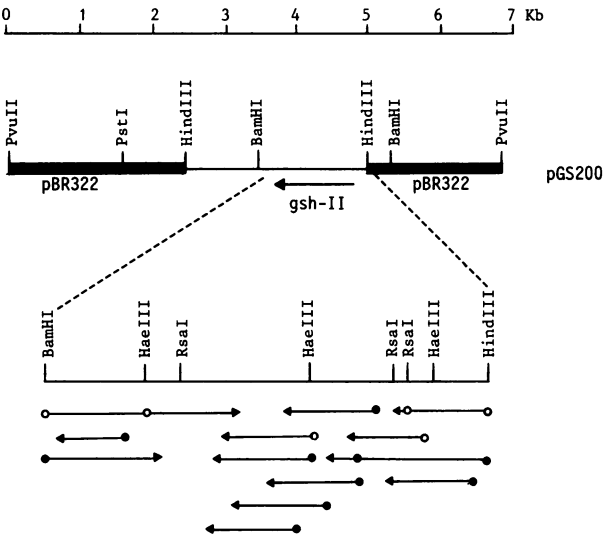

Figure 1. Physical map of the recombinant plasmid pGS200 (Murata et al.,1983) and the strategy for the DNA sequencing of gsh-II gene. The restriction sites of pGS200 DNA used for sequencing indicated at the map coordinated on the top column. At the bottom column in the expanding scale, the extent and direction of DNA sequencing are indicated by horizontal arrows. The open and closed circles represent the 5' termini of the cloned fragments generated by restriction enzyme digestion and sonication, respectively.

digested pGS200 with BamHI restriction endonuclease and cloned into the BamHI site of pBR325. These plasmids were used for transformation of E.coli B C1001 (gsh-II$^-$). The resulting transformants (gsh-II$^+$) were selected according to the method of Murata et al. (2). The restriction mapping has indicated that the gsh-II gene is located within the HindIII-BamHI fragment (1400 bp) of pGS200 (Fig. 1).

Nucleotide sequence of the gsh-II gene

To determine the nucleotide sequence of the gsh-II gene, we first digested pGS200 with the combination of BamHI and HindIII to isolate the 1400 bp DNA fragment. The fragment was dissected further by either restriction enzymes or sonication and inserted into the corresponding cloning sites of the M13 vector. After transfection in E.coli JM101 strain, the recombinant phages were propagated and the resulting phage DNA was submitted to sequencing (Fig.1).

The sequence of the 1,478 bp chromosomal DNA containing the gsh-II gene is shown in Figure 2. Examination of the nucleotide sequence shows only one

```
  1                                                                  AAGCT

  6 TCAGCAGTGGCAGAAGATTGCAATTGCTGCCTGTGAGCAGTGTGGTCGTAACCGGGTGCC

 66 GAAATCCGTCCAGCGATGGATCTGGAAGCCTGGTGTGCAGAGCAGGATGAAGGACTGAAA

126 CTGAATCTTCACCCGCGCGCCAGTAACAGCATCAATACGTTGCCGTTACCGGTTGAACGC

186 GTCCGCCTGCTGATTGGCCCGGAAGGCGGTTTATCGGCAGATGAAATTGCCATGACTGCC
                                                      -35 SQ
246 CGCTATCAATTTACTGATATCCTGTTGGGACCTCGCGTTTTGCGTACAGAGACAACTGCG
              Pribnow          start of mRNA
306 CTCACCGCCATTACCGCGCTACAAGTACGATTTGGCGATTTGGGCTAACGGAGAAGAATA
                                                        S D
366 ATG ATC AAG CTC GGC ATC GTG ATG GAC CCC ATC GCA AAC ATC AAC
  1 Met Ile Lys Leu Gly Ile Val Met Asp Pro Ile Ala Asn Ile Asn

411 ATC AAG AAA GAT TCC AGT TTT GCT ATG TTG CTG GAA GCA CAG CGT
 16 Ile Lys Lys Asp Ser Ser Phe Ala Met Leu Leu Glu Ala Gln Arg

456 CGT GGT TAC GAA CTT CAC TAT ATG GAG ATG GGC GAT CTG TAT CTG
 31 Arg Gly Tyr Glu Leu His Tyr Met Glu Met Gly Asp Leu Tyr Leu

501 ATC AAT GGT GAA GCC CGC GCC CAT ACC CGC ACG CTG AAC GTG AAG
 46 Ile Asn Gly Glu Ala Arg Ala His Thr Arg Thr Leu Asn Val Lys

546 CAG AAC TAC GAA GAG TGG TTT TCG TTC GTC GGT GAA CAG GAT CTG
 61 Gln Asn Tyr Glu Glu Trp Phe Ser Phe Val Gly Glu Gln Asp Leu

591 CCG CTG GCC GAT CTC GAT GTG ATC CTG ATG CGT AAA GAC CCG CCG
 76 Pro Leu Ala Asp Leu Asp Val Ile Leu Met Arg Lys Asp Pro Pro

636 TTT GAT ACC GAG TTT ATC TAC GCG ACC TAT ATT CTG GAA CGT GCC
 91 Phe Asp Thr Glu Phe Ile Tyr Ala Thr Tyr Ile Leu Glu Arg Ala

681 GAA GAG AAA GGG ACG CTG ATC GTT AAC AAG CCG CAG AGC CTG CGC
106 Glu Glu Lys Gly Thr Leu Ile Val Asn Lys Pro Gln Ser Leu Arg

726 GAC TGT AAC GAG AAA CTG TTT ACC GCC TGG TTC TCT GAC TTA ACG
121 Asp Cys Asn Glu Lys Leu Phe Thr Ala Trp Phe Ser Asp Leu Thr

771 CCA GAA ACG CTG GTT ACG CGC AAT AAA GCG CAG CTA AAA GCG TTC
136 Pro Glu Thr Leu Val Thr Arg Asn Lys Ala Gln Leu Lys Ala Phe

816 TGG GAG AAA CAC AGC GAC ATC ATT CTT AAG CCG CTG GAC GGT ATG
151 Trp Glu Lys His Ser Asp Ile Ile Leu Lys Pro Leu Asp Gly Met

861 GGC GGC GCG TCG ATT TTC CGC GTG AAA GAA GGC GAT CCA AAC CTC
166 Gly Gly Ala Ser Ile Phe Arg Val Lys Glu Gly Asp Pro Asn Leu

906 GGC GTG ATT GCC GAA ACC CTG ACT GAG CAT GGC ACT CGC TAC TGC
181 Gly Val Ile Ala Glu Thr Leu Thr Glu His Gly Thr Arg Tyr Cys

951 ATG GCG CAA AAT TAC CTG CCA GCC ATT AAA GAT GGC GAC AAA CGC
196 Met Ala Gln Asn Tyr Leu Pro Ala Ile Lys Asp Gly Asp Lys Arg
```

```
 996 GTG CTG GTG GTG GAT GGC GAG CCG GTA CCG TAC TGC CTG GCG CGT
 211 Val Leu Val Val Asp Gly Glu Pro Val Pro Tyr Cys Leu Ala Arg

1041 ATT CCG CAG GGG GGC GAA ACC CGT GGC AAT CTG GCT GCC GGT GGT
 226 Ile Pro Gln Gly Gly Glu Thr Arg Gly Asn Leu Ala Ala Gly Gly

1086 CGC GGT GAA CCT CGT CCG CTG ACG GAA AGT GAC TGG AAA ATC GCC
 241 Arg Gly Glu Pro Arg Pro Leu Thr Glu Ser Asp Trp Lys Ile Ala

1131 CGT CAG ATC GGG CCG ACG CTG AAA GAA AAA GGG CTG ATT TTT GTT
 256 Arg Gln Ile Gly Pro Thr Leu Lys Glu Lys Gly Leu Ile Phe Val

1176 GGT CTG GAT ATC ATC GGC GAC CGT CTG ACT GAA ATT AAC GTC ACC
 271 Gly Leu Asp Ile Ile Gly Asp Arg Leu Thr Glu Ile Asn Val Thr

1221 AGC CCA ACC TGT ATT CGT GAG ATT GAA GCA GAG TTT CCG GTG TCG
 286 Ser Pro Thr Cys Ile Arg Glu Ile Glu Ala Glu Phe Pro Val Ser

1266 ATC ACC GGA ATG TTA ATG GAT GCC ATC GAA GCA CGT TTA CAG CAG
 301 Ile Thr Gly Met Leu Met Asp Ala Ile Glu Ala Arg Leu Gln Gln

1311 CAG TAA CCCACCTTAGCGAGAAGGATCTCGTTGAGACTCTGAGTGACAGCGCCCTTCT
 316 Gln Ter                                                 I R

1369 TTCCACGCATACTGGGCGCTGTTGCTTTTTTGAACCAGGAAACAGAACCTCTGACAATGA
                I R          termination of mRNA
1429 TGAATTTACAGCATCACTTTCTTATTGCCATGCCTGCTCTCCAGGATCC
```

Figure 2. Complete nucleotide sequence of gsh-II gene and the flanking regulatory unit sequences.
The nucleotide sequence of the gene is indicated and nucleotides are numbered from the HindIII site. The amino acid sequence of GSH-II predicted from the equences is given below the sequence. Several regulatory sequences flanking the gsh-II gene are indicated with underlines. They include the Shine-Dalgarno (SD) sequences, the Pribnow box and the conserved sequence (-35 SQ) located at about -10 and -35 nucleotides, respectively, upstream from the starting point of mRNA synthesis which was tentatively assigned to position 268 or 267 as judged from the topology of the reguratory signal sequence described above. Downstream from the presumed terminaton codon of the gsh-II gene at positions 1314-1316, the inverted repeat sequences (IR) are presented at positions 1356-1364 and 1382-1390, followed by 6 successive T's which is a preferable site for the termination of mRNA synthesis.

possible open reading frame, starting at position 364 and terminating at position 1314, which was sufficiently long enough to code for a polypeptide of about 38 kd in molecular weight. In order to ascertain the initiation site of the gsh-II, we determined the amino terminal sequence of purified GSH-II. They were Met-Ileu-Lys-Leu-Gly-Ileu-Val-Met-Asp-Pro-Ileu-Ala-Asn-Ileu-Asn-Ileu-Lys-Lys-Asn-Ser-X-Phe-Ala-Met-Leu-Leu-Glu (Table 1). This sequence completely agrees with that of the first 27 amino acid predicted from the DNA sequence. This open reading region can code for a polypeptide of 316 amino acids. The amino acid sequence predicted for the GSH-II is

Table 1. Coincidence of the amino acid sequence at the NH_2-terminus of GSH-II with that predicted from nucleotide sequence.

NH_2-terminal Residue Number	Codon	Amino Acid Residue	NH_2-terminal Residue Number	Codon	Amino Acid Residue
1	ATG	Met	15	AAC	Asn
2	ATC	Ile	16	ATC	Ile
3	AAG	Lys	17	AAG	Lys
4	CTC	Leu	18	AAA	Lys
5	GGC	Gly	19	GAT	Asp
6	ATC	Ile	20	TCC	Ser
7	GTG	Val	21	AGT	X
8	ATG	Met	22	TTT	Phe
9	GAC	Asp	23	GCT	Ala
10	CCC	Pro	24	ATG	Met
11	ATC	Ile	25	TTG	Leu
12	GCA	Ala	26	CTG	Leu
13	AAC	Asn	27	GAA	Glu
14	ATC	Ile			

X indicates unidentified amino acid

shown in Fig.2. The amino acid composition to be as follows : Ala_{22}, Arg_{19}, Asn_{12}, Asp_{20}, Cys_4, Gln_{11}, Glu_{27}, Gly_{25}, His_4, $Ileu_{26}$, Leu_{33}, Lys_{18}, Met_{10}, Phe_{11}, Pro_{17}, Ser_{10}, Thr_{19}, Trp_4, Tyr_9, and Val_{15}. The molecular weight calculated from the predicted amino acid squence is 35,559 daltons, which agrees well with the molecular weight of the GSH subunit (38 kd) estimated by SDS-PAGE (5). Since an apparent molecular weight of the GSH-II in an the active state estimated by molecular sieve chromatography is about 152 kd, active GSH-II seems to be a tetrameric enzyme composed of four identical subunits. Codon usage for E.coli GSH-II derived from DNA sequence data is shown in Table 2.

Transcriptional signals

Prokaryotic consensus sequences for transcriptional initiation have well been documented (10-12). In the precise DNA sequence shown in Figure 2, we find two hexanucleotides TTGCCA and TTTACT at positions 232 and 255, respectively, preceding to the initiation codon ATG at position 366. The former sequence matches in five out of six positions to the consensus

Table 2.Codon usage in the E.coli B GSH-II gene

Phe	UUU	7	Tyr	UAU	3
	UUC	4		UAC	6
Leu	UUA	3	End	UAA	1
	UUG	1	End	UAG	0
Leu	CUU	2	HIs	CAU	2
	CUC	3		CAC	2
	CUA	1	Gln	CAA	1
	CUG	23		CAG	10
Ile	AUU	10	Asn	AAU	4
	AUC	16		AAC	8
	AUA	0	Lys	AAA	13
Met	AUG	10		AAG	5
Val	GUU	3	Asp	GAU	11
	GUC	2		GAC	9
	GUA	1	Glu	GAA	17
	GUG	9		GAG	10
Ser	UCU	1	Cys	UGU	2
	UCC	1		UGC	2
	UCA	0	End	UGA	0
	UCG	3	Trp	UGG	4
Pro	CCU	1	Arg	CGU	11
	CCC	1		CGC	8
	CCA	4		CGA	0
	CCG	11		CGG	0
Tyr	ACU	3	Ser	AGU	2
	ACC	9		AGC	3
	ACA	0	Arg	AGA	0
	ACG	7		AGG	0
Ala	GCU	2	Gly	GGU	8
	GCC	10		GGC	12
	GCA	4		GGA	1
	GCG	6		GGG	4

sequence TTGACA conserved in the "-35 region" upstream from the initiation site (10) and is identical to that reported for purf gene (13). The latter sequence also agrees in four out of six nucleotides with the Pribnow box TATAAT (10). Furthermore, these consensus sequences are separated from each other by 17 bp, presumably being the most preferrable spacing for an effcient transcription (11). Regarding a transcriptional termination signal, it has not been studied so extensively. The inverted repeat sequences can be located at positions 1356 and 1382, which can form a stable hairpin loop structure. This inverted repeats are immediately followed by T-rich sequence. This sequence arrangemes is ofen found in the prokaryotic terminal region of mRNA (10).

DISCUSSION

We have established the complete nucleotide sequence of E.coli B gsh-II gene, including the flanking regulatory regions. The DNA sequence upstream from the 5'-terminus of the gsh-II gene (position 220-270) is slightly A-T rich, where the promoter is expected to be located. The sequence of putative "-35 region" and Pribnow boxes of the gsh-II gene are both very similar to the known consensus nucleotide sequences (10). In this region, another Pribnow box like sequence TATCCT is found at position 264. Hoever the sequence is separated from the "-35 region" by 25 bp, the spacing being far from the typical spacing of 17 nucleotides (10). Importance of the critical spacing between the two conserved sequence is further strengthened by the DNA of considering the direct contact sites of RNA polymerase with the promoter region on three-dimensional model (12). From these points, we assigned the sequence TTTACT at position 255 to be the Pribnow box. However, further biochemical and genetic experiments will be required to confirm our tentative promoter sequence as well as the transcriptionl start point. Furthermore, about 10 nucleotides upstream from the initiation codon ATG a potential Shine-Dalgarno sequence GGAG (15) is located (Fig. 2).

About 40 bp downstream from the translational termination codon TAA, there exist GC-rich inverted repeated sequences followed by a streatch of successive Ts. This can form a very stable structure with a G of -23.kcal/mol, calculated according to the method of Tinoco et al (14). Presumably they constitute the transcriptional termination signal.

A molecular weight calculated from the 316 amino acid residues encoded from the gsh-II gene is 35,549 daltons, and is in good agreement with that estimated from SDS-PAGE (38 kd). Recently, by the high performance liquid chromatography using a SW 3000 column (Toyo Soda), a more precise molecular weight of 140 kd for the GSH-II holoenzyme was determined. As the GSH-II is comprized of four identical subunits (5), the subunit molecular weight (35.6 kd) caluculated from the DNA sequence data fits even better to this estimate than does the value (38 kd) by SDS-PAGE. Codon usage in the gsh-II gene shows no strong bias from the common tendancy (Table 2). We can estimate the frequency of optimal codon usage (Fop) to be 0.72, according to the method of Ikemura and Ozeki (16). Thus, gsh-II gene seems to be only moderately expressed in E.coli cells.

We have been establishing a bioreacter system for the production of glutathione by the use of the immobilized E.coli cells modified by genetic engineering teqniques (1-4). When both gsh-I and gsh-II genes inserted

simultaneously into pBR325 were introduced into the E.coli K-12 C600 cells, the productivity of glutathione was very much enhanced. While the gsh-I gene is now being sequenced, the knowledge of nucleotide sequences of both genes will enable us to improve the expression rate of these genes by gene manipulation.

ACKNOWLEDGMENTS

We thank Prof. M. Takanami, Institute for Chemical Research of Kyoto University, Assoc. Prof. M. Sakaguchi, Dr. K. Murata, Dr. Y. Fukuda, Mr. K. Watanabe, Research Institute for Food Science, Kyoto University, and Mr. T. Miya, Reseach Laboratory of Kojin Company, Ltd., for many helpful discussions.

REFERENCES

1. Murata, K. and Kimura, A. (1982) Appl. Environ. Microbiol. 44, 1444-1448.
2. Murata, K., Miya, T., Gushima, H. and Kimura, A. (1983) Agr. Biol. Chem. 47, 1381-1383.
3. Gushima, H., Miya, T., Murata, K. and Kimura, A. (1983) Agr. Biol. Chem. 47, 1927-1928.
4. Gushima, H., Miya, T., Murata, K. and Kimura, A. (1983) J. Appl. Biochem. 5, 43-52.
5. Gushima, H., Miya, T., Murata, K. and Kimura, A. (1983) J. Appl. Biochem. 5, 210-2186.
6 Murata, K., Tani, K., Kato, J. and Chibata, I. (1981) Agr. Biol. Chem. 45, 2131-2132.
7. Messing, J. and Viera, J. (1982) Gene. 19, 269-272
8. Deininger, P.L. (1983) Anal. Biochem. 129. 216-223.
9. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad.Sci. USA. 74, 5463-5467.
10. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353.
11. Hawley, D.K. and McClure, W.R. (1983) Nucl. Acids Res. 11, 2237-2255.
12. Siebenlist, U., Simpson, R.B. and Gilbert, W. (1980) Cell 20, 269-281.
13. Tso, J.Y., Zalkin, T.J.Y., H.van Cleemput, H.V., Yanofsky, C. and Smith J.M. (1982) J. Biol. Chem. 257, 3525-3531.
14. Tinoco, I.Jr., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) Nat. New Biol. 246, 40-41.
15. Shine, J. and Dalgarno, L. (1975) Proc. Natl. Acad. Sci. USA. 72, 784-788.
16. Ikemura, T. and Ozeki, H. (1981) J. Mol. Biol. 146, 1-21.