JOURNAL OF BACTERIOLOGY, Sept. 1990, p. 5497–5500
0021-9193/90/095497-04$02.00/0

Vol. 172, No. 9

The Complete Nucleotide Sequence of the *Escherichia coli* Gene *appA* Reveals Significant Homology between pH 2.5 Acid Phosphatase and Glucose-1-Phosphatase

JANIE DASSA, CHRISTIAN MARCK, AND PAUL L. BOQUET*

Service de Biochimie, Département de Biologie, CEN Saclay 91191 Gif-sur-Yvette, Cedex, France

Received 28 December 1989/Accepted 26 June 1990

The whole nucleotide sequence of *Escherichia coli* gene *appA*, which encodes periplasmic phosphoanhydride phosphohydrolase (optimum pH, 2.5), and its flanking regions was determined. The AppA protein is significantly homologous to the product of the nearby gene *agp*, acid glucose-1-phosphatase. Because identical amino acids are distributed over the whole lengths of the proteins, it is likely that *appA* and *agp* originate from the same ancestor gene.

Gene *appA* lies near minute 22 on the *Escherichia coli* linkage map and encodes a periplasmic acid phosphatase of unknown function (2, 4–6, 13, 18). A nucleotide sequence corresponding to a 5′-terminal portion of *appA* and to a small upstream region has already been published by Touati and Danchin (16). Here we present the complete nucleotide sequence of *appA* and its flanking regions. This sequence shows marked differences from that already reported (16). The deduced *appA* amino acid sequence shows very significant homology to that of the product of another *E. coli* gene, *agp*, which encodes periplasmic glucose-1-phosphatase, an enzyme that scavenges glucose from external glucose-1-phosphate and was previously studied in our laboratory (10–12).

A physical map of the *appA* region on recombinant plasmid pPB1132, previously described (2), is given in Fig. 1, along with the strategy used for sequencing. The nucleotide sequence of a DNA fragment known to contain the whole *appA* gene, on the basis of previous deletion and transposon insertion analysis (2), is shown in Fig. 2. This sequence contains a large open reading frame (ORF) starting 4 nucleotides downstream from the unique *Cla*I restriction site and extends over 1,298 nucleotides. Its location, direction of transcription, and size are in perfect agreement with previously reported data on the physical mapping of *appA* deduced from analysis of several *appA-phoA* hybrid proteins obtained by insertion of Tn*phoA* into *appA* (2). The nucleotide sequence shown is numbered starting from a position which corresponds to the termination codon (TAA) of a large upstream ORF (ORFA) which is separated from *appA* by a region of 185 base pairs (the nucleotide sequence of ORFA will be reported elsewhere). The region separating ORFA from *appA* contains a very short open reading frame (ORFX), which encodes a peptide of 30 amino acids but contains no typical transcription termination signal. It does, however, show an intergenic palindromic sequence (7) just after ORFX, between positions 118 and 150. The region upstream of *appA* contains sequences of hexanucleotides which, although not typical, resemble −35 and −10 promoter sequences (TTAGCA and AATAAT, respectively). A Shine-Dalgarno-like sequence (AAGCGG) is found at a reasonable distance from the ATG initiation codon of *appA*. At a short distance downstream from the *appA* termination codon lies a typical Rho-independent transcription termination sequence which is oriented in a direction opposite to that of *appA*. No other ORF oriented like *appA* was found in the region lying between *appA* and the *Pvu*I site at the end of the sequence shown.

The amino acid sequence corresponding to *appA* (Fig. 2) shows the presence of a 22-amino-acid N-terminal hydrophobic signal peptide with lysine as a positively charged amino acid in position 2 (19). The putative cleavage site by the leader peptidase is located between the recognition sequence Ala–Phe–Ala-22 and Gln-23. The N-terminal amino acid sequence of the mature acid phosphatase was determined after purification of the enzyme to homogeneity as previously described (6). The first seven amino acids of the N-terminal part of the purified protein was identified by using an Applied Biosystems 477A gas phase sequenator and on-line automated high-pressure liquid chromatography. The sequence found (NH_2-Gln-Ser-Glu-Pro...) confirms the proposed position for cleavage from the preprotein. Consequently, mature pH 2.5 acid phosphatase is 410 amino acids long and has an M_r of 44,644 compared with the 45,000 previously estimated (6).

The amino acid sequence of the product of *appA* was compared with those of all of the sequenced proteins available in version 61 of the GenPro data bank, which includes several phosphatases of eucaryotic and procaryotic origins. Apart from the short Arg-His-Gly sequence (positions 38 to 40) previously shown by Bazan et al. to exist in a similar location in several acid phosphatases and phosphoglucomutates of eucaryotic origin and proposed to belong to the phosphatase active site of the enzymes (1), no significant homologies could be found, even with other *E. coli* periplasmic phosphatases, such as the products of *phoA* (alkaline phosphatase), *ushA* (5′-nucleotidase), or *cpdB* (2′-3′-cyclic phosphodiesterase). Surprisingly, however, we disclosed clear homology between the sequence of AppA and that of glucose-1-phosphatase, the product of the nearby gene *agp* (10–12). Although the Agp protein is shorter by 11 amino acid residues than the product of *appA*, isolated identical amino acids or short stretches of identical or structurally equivalent amino acids were found over the whole length of the polypeptides (Fig. 3). Dot matrix analysis (data not shown) revealed four regions of the two proteins exhibiting 36% homology. Altogether, these four sequences represent

* Corresponding author.

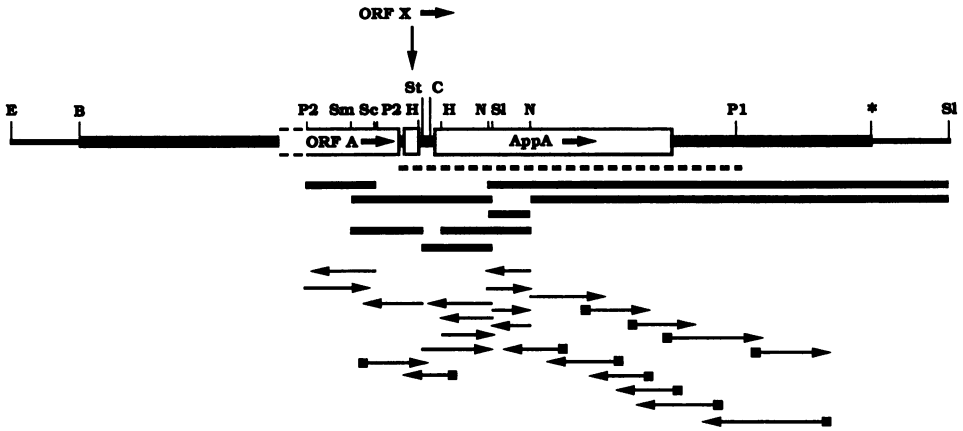

FIG. 1. Restriction map of the *appA* region on plasmid pPB1132, subcloning, and strategy for DNA sequencing. The position of the *appA* gene (open box) on the *Bam*HI-*Bgl*II DNA fragment (heavy bar) previously cloned into the *Bam*HI site of pBR322 (thin bar) and its direction of transcription are derived from previous deletion-and-insertion analysis (3). The DNA fragments subcloned into M13 derivatives (20) are indicated below as heavy bars. Arrowed segments show the length and direction of the nucleotide sequence determined in each subclone. Boxes at the origins of some of the segments indicate the use of a specific oligonucleotide as a primer for the sequencing reactions which were performed as described in references 14 and 15. Abbreviations: B, *Bam*HI; C, *Cla*I; E, *Eco*RI; H, *Hpa*I; N, *Nru*I; P1, *Pvu*I; P2, *Pvu*II; Sc, *Sca*I; Sl, *Sal*I; Sm, *Sma*I; St, *Stu*I. The sequence reported in Fig. 2 is illustrated by the broken line.

68 and 70% of the total protein sequence (including the signal peptide) for AppA and Agp, respectively. Furthermore, six of the seven cysteine residues of AppA lie in the same position as or in a position extremely close to that found in the Agp protein. AppA has seven tryptophan residues, five of which are conserved in Agp. When the tyrosine and phenylalanine residues, with a Tyr-to-Phe substitution considered acceptable, are taken together, relative homology is found for these aromatic amino acid residues. The homology is also well conserved for proline residues. Moreover, the hydropathic profiles of the two proteins, obtained as described by Kyte and Doolittle (8), matched closely (data not shown).

This presence of isolated or clustered identical amino acids in several regions distributed over the whole proteins argues in favor of the divergence of the two enzymes from a

```
  1  *           M  W  Y  L  L  W  F  V  G  I  L  L  M  C  S  L  S  T  L  V  L  V  W  L  D  P  R  L  K  S  *        ---
  1  TAAggagcagaaacaATGTGGTATTTACTTTGGTTCGTCGGCATTTTGTTGATGTGTTCGCTCTCCACCCTTGTGTTGGTATGGCTGGACCCGCGTCTGAAAAGTTAAcgaacgtaggcc  120

  1  -------    -35    ----------    -10                SD              M  K  A  I  L  I  P  F  L  S  L  L  I  P  L  T  P  Q   18
121  tgatgcggcgcattagcatcgcatcaggcaatcaataatgtcagatatgaaaagcggaaacatatcgATGAAAGCGATCTTAATCCCATTTTTATCTCTTCTGATTCCGTTAACCCCGCA  240

 19  S  A  F  A  Q  S  E  P  E  L  K  L  E  S  V  V  I  V  S  R  H  G  V  R  A  P  T  K  A  T  Q  L  M  Q  D  V  T  P  D  A   58
241  ATCTGCATTCGCTCAGAGTGAGCCGGAGCTGAAGCTGGAAAGTGTGGTGATTGTCAGTCGTCATGGTGTGCGTGCTCCAACCAAGGCCACGCAACTGATGCAGGATGTCACCCCAGACGC  360

 59  W  P  T  W  P  V  K  L  G  W  L  T  P  R  G  G  E  L  I  A  Y  L  G  H  Y  Q  R  Q  R  L  V  A  D  G  L  L  A  K  K  G   98
361  ATGGCCAACCTGGCCGGTAAAACTGGGTTGGCTGACACCGCG GGTGGTGAGCTAATCGCCTATCTCGGACATTACCAACGCCAGCGTCTGGTAGCCGACGGATTGCTGGCGAAAAAGGG  480

 99  C  P  Q  S  G  Q  V  A  I  I  A  D  V  D  E  R  T  R  K  T  G  E  A  F  A  A  G  L  A  P  D  C  A  I  T  V  H  T  Q  A  138
481  CTGCCCGCAGTCTGGTCAGGTCGCGATTATTGCTGATGTCGACGAGCGTACCCGTAAAACAGGCGAAGCCTTCGCCGCCGGGCTGGCACCTGACTGTGCAATAACCGTACATACCCAGGC  600

139  D  T  S  S  P  D  P  L  F  N  P  L  K  T  G  V  C  Q  L  D  N  A  N  V  T  D  A  I  L  S  R  A  G  G  S  I  A  D  F  T  178
601  AGATACGTCCAGTCCCGATCCGTTATTTAATCCTCTAAAAACTGGCGTTTGCCAACTGGATAACGCGAACGTGACTGACGCGATCCTCAGCAGGGCAGGAGGGTCAATTGCTGACTTTAC  720

179  G  H  R  Q  T  A  F  R  E  L  E  R  V  L  N  F  P  Q  S  N  L  C  L  K  R  E  K  Q  D  E  S  C  S  L  T  Q  A  L  P  S  218
721  CGGGCATCGGCAAACGGCGTTTCGCGAACTGGAACGGGTGCTTAATTTTCCGCAATCAAACTTGTGCCTTAAACGTGAGAAACAGGACGAAAGCTGTTCATTAACGCAGGCATTACCATC  840

219  E  L  K  V  S  A  D  N  V  S  L  T  G  A  V  S  L  A  S  M  L  T  E  I  F  L  L  Q  Q  A  Q  G  M  P  E  P  G  W  G  R  258
841  GGAACTCAAGGTGAGCGCCGACAATGTCTCATTAACCGGTGCGGTAAGCCTCGCATCAATGCTGACGGAGATATTTCTCCTGCAACAAGCACAGGGAATGCCGGAGCCGGGGTGGGGAAG  960

259  I  T  D  S  H  Q  W  N  T  L  L  S  L  H  N  A  Q  F  Y  L  L  Q  R  T  P  E  V  A  R  S  R  A  T  P  L  L  D  L  I  K  298
961  GATCACCGATTCACACCAGTGGAACACCTTGCTAAGTTTGCATAACGCGCAATTTTATTTGCTACAACGCACGCCAGAGGTTGCCCGCAGCCGCGCCACCCCGTTATTAGATTTGATCAA 1080

299  T  A  L  T  P  H  P  P  Q  K  Q  A  Y  G  V  T  L  P  T  S  V  L  F  I  A  G  H  D  T  N  L  A  N  L  G  G  A  L  E  L  338
1081 GACAGCGTTGACGCCCCATCCACCGCAAAAACAGGCGTATGGTGTGACATTACCCACTTCAGTGCTGTTTATCGCCGGACACGATACTAATCTGGCAAATCTCGGCGGCGCACTGGAGCT 1200

339  N  W  T  L  P  G  Q  P  D  N  T  P  P  G  G  E  L  V  F  E  R  W  R  R  L  S  D  N  S  Q  W  I  Q  V  S  L  V  F  Q  T  378
1201 CAACTGGACGCTTCCCGGTCAGCCGGATAACACGCCGCCAGGTGGTGAACTGGTGTTTGAACGCTGGCGTCGGCTAAGCGATAACAGCCAGTGGATTCAGGTTTCGCTGGTCTTCCAGAC 1320

379  L  Q  Q  M  R  D  K  T  P  L  S  L  N  T  P  P  G  E  V  K  L  T  L  A  G  C  E  E  R  N  A  Q  G  M  C  S  L  A  G  F  418
1321 TTTACAGCAGATGCGTGATAAAACGCCGCTGTCATTAAATACGCCGCCCGGAGAGGTGAAACTGACCCTGGCAGGATGTGAAGAGCGAAATGCGCAGGGCATGTGTTCGTTGGCAGGTTT 1440

419  T  Q  I  V  N  E  A  R  I  P  A  C  S  L  *          ----------    ----------                                        433
1441 TACGCAAATCGTGAATGAAGCACGCATACCGGCGTGCAGTTTGTAAtgcataaaaaagagcattcagttacctgaatgctctgaggctgatgacaaacgaagaactgtctaatgcgtaga 1560

1561 ccggaaaaggcgttcacgccgcatccggccactttcagtttt::ctctttctcggagtaactataaccgtaatagttatagccgtaactgtaagcggtgctggcgcgtttaatcacaccat 1680

1681 tgaggatagcgcctttaatattgacgcctgcctgttccagacgctgcattgacaaactcacctctttggcggtgttcaagccaaaacgcgcaaccagcaggctggtgccaacagaacgcc 1800
                                                                                                                  Pvu I
1801 ccacgaccgcggcatcactcaccgccagcatcggcggcgtatcgacaatcaccagatcgtaatggtcgttcgcccattccagtaattgacgcatccgatcg
```

FIG. 2. Nucleotide sequence of the *appA* gene region of *E. coli*. The nucleotide sequence is numbered starting from the termination codon of ORFA and extends to the unique *Pvu*I site of pPB1132. The derived amino acid sequence is indicated for both ORFX and *appA*. Regions of dyad symmetry are indicated with dashes. Putative Shine-Dalgarno sequences are indicated (SD), as are −35 and −10 promoterlike sequences for *appA*. Sequence analysis was performed by using DNA Strider software (9). The asterisks indicate the termination codons of ORFA (nucleotides 1 to 3) and ORFX (nucleotides 106 to 108).

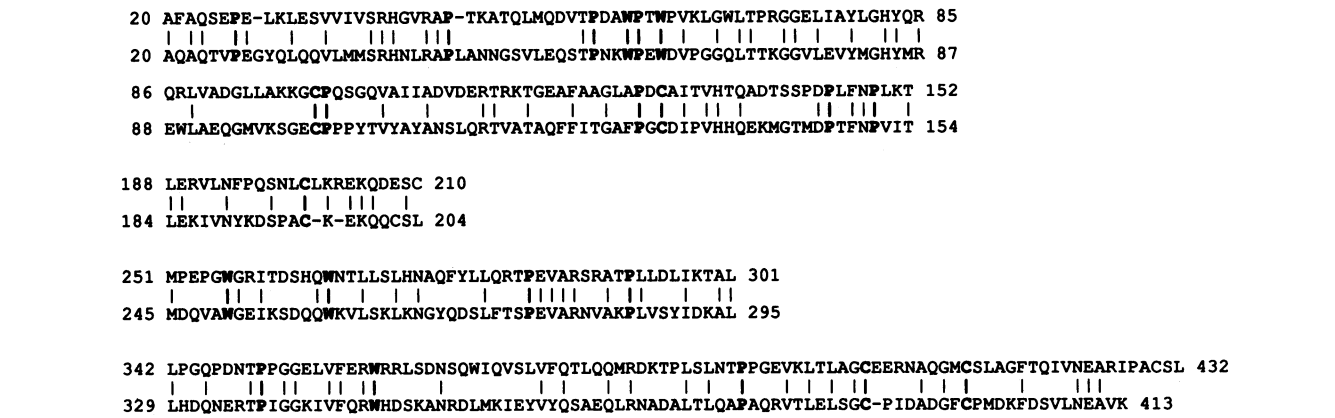

FIG. 3. Comparison of the deduced amino acid sequences of AppA and Agp. The four most conserved regions between the sequences of AppA (top lines) and Agp (bottom lines) shown were identified by dot matrix analysis. Conserved cysteine (C), tryptophan (W), and proline (P) residues are indicated in boldface.

common ancestor gene. However, the Asp-Ser-Ala-Ala sequence, which is found in Agp at positions 157 to 160 (12) and also exists in the active site of alkaline phosphatase and those of other typical serine hydrolases (3), is not present on AppA. Moreover, Agp has an Arg-His-Asn sequence in place of the Arg-His-Gly putative active site of AppA. If the gene duplication hypothesis is correct and if the positions of the respective active sites on the two proteins are further confirmed, it would be interesting to identify the few changes that lead to differences in both the substrate preference and the mechanism of reaction of the two related enzymes.

Expression of *appA* on the chromosome is highly stimulated by entry of the bacteria into the stationary phase of growth, by anaerobiosis, and by inorganic phosphate starvation (18). It also depends on the allelic state of the gene *appR* (17, 18). Expression of *agp*, by contrast, is not influenced by any of these factors (11). The *agp* gene and its promoter lie between transcription termination signals (12), but *appA* transcription can proceed over upstream genes from an exogenous promoter on a recombinant plasmid (2, 18). Accordingly, the nucleotide sequence between ORFA and *appA* contains no transcription termination signal and no significant homology was found between the nucleotide sequences preceding *appA* and *agp* (C. Marck et al., unpublished data). The presence of a typical palindromic unit upstream of *appA* might be an indication of the existence of a chromosomal rearrangement in this region (7). There is no direct evidence that the −10 and −35 sequences found immediately upstream of *appA* constitute the main promoter of this gene on the chromosome. The existence of a large ORF (ORFA) upstream of *appA* is consistent with the previously reported expression of alkaline phosphatase from a Tn*phoA* transposon inserted in the same region (2). This suggests that *appA* belongs to a polycistronic operon that specifies at least another extracytoplasmic protein encoded by ORFA.

Homology searches in the GenPro data bank were performed by using the CITI2 computing facilities on a VAX 8530 computer with the help of the French Ministère de la Recherche et de la Technologie (programme mobilisateur Essor de Biotechnologies).

We are grateful to J. M. Buhler and C. Doira for preparing the oligonucleotides used in sequencing *appA*.

LITERATURE CITED

1. **Bazan, J. F., R. J. Fletterick, and S. J. Pilkis.** 1989. Evolution of a bifunctional enzyme: 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase. Proc. Natl. Acad. Sci. USA **86**:9642–9646.
2. **Boquet, P. L., C. Manoil, and J. Beckwith.** 1987. Use of Tn*phoA* to detect genes for exported proteins in *Escherichia coli*: identification of the plasmid-encoded gene for periplasmic acid phosphatase. J. Bacteriol. **169**:1663–1669.
3. **Bradshaw, R. A., F. Cancedda, L. H. Ericsson, P. A. Neumann, S. P. Piccoli, M. J. Schlesinger, S. Shrieffer, and K. A. Walsh.** 1981. Amino acid sequence of *Escherichia coli* alkaline phosphatase. Proc. Natl. Acad. Sci. USA **78**:3473–3477.
4. **Dassa, E., and P. L. Boquet.** 1981. Is the acid phosphatase of *Escherichia coli* with pH optimum of 2.5 a polyphosphate depolymerase? FEBS Lett. **135**:148–150.
5. **Dassa, E., and P. L. Boquet.** 1985. Identification of the gene *appA* for the acid phosphatase (pH optimum 2.5) of *Escherichia coli*. Mol. Gen. Genet. **200**:68–73.
6. **Dassa, E., M. Cahu, B. Desjoyaux-Cherel, and P. L. Boquet.** 1982. The acid phosphatase with optimum pH of 2.5 of *Escherichia coli*: physiological and biochemical study. J. Biol. Chem. **257**:6669–6676.
7. **Gilson, E., J. M. Clément, D. Perrin, and M. Hofnung.** 1987. Palindromic units: a case of highly repetitive DNA sequences in bacteria. Trends Genet. **3**:226–230.
8. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **167**:105–132.
9. **Marck, C.** 1988. 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Res. **16**:1829–1836.
10. **Pradel, E., and P. L. Boquet.** 1988. Acid phosphatases of *Escherichia coli*: molecular cloning and analysis of *agp*, the structural gene for a periplasmic acid glucose phosphatase. J. Bacteriol. **170**:4916–4923.
11. **Pradel, E., and P. L. Boquet.** 1989. Mapping of the *Escherichia coli* acid glucose-1-phosphatase gene *agp* and analysis of its expression in vivo by use of an *agp-phoA* protein fusion. J. Bacteriol. **171**:3511–3517.
12. **Pradel, E., C. Marck, and P. L. Boquet.** 1990. Nucleotide sequence and transcriptional analysis of the *Escherichia coli agp* gene encoding periplasmic acid glucose-1-phosphatase. J. Bacteriol. **172**:802–807.
13. **Rao, N. N., M. F. Roberts, and A. Torriani.** 1987. Polyphosphate accumulation and metabolism in *Escherichia coli*, p. 213–219. *In* A. Torriani-Gorini, F. G. Rothman, S. Silver, A. Wright, and E. Yagil (ed.), Phosphate metabolism and cellular regulation in microorganisms. American Society for Microbiology, Washington, D.C.
14. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
15. **Tabor, S., and C. C. Richardson.** 1987. DNA sequence analysis with a modified bacteriophage T7 RNA polymerase. Proc. Natl. Acad. Sci. USA **84**:4767–4771.

16. **Touati, E., and A. Danchin.** 1987. The structure of the promoter and amino terminal region of the pH 2.5 acid phosphatase structural gene, *appA*, of *E. coli*: a negative control of transcription mediated by cyclic AMP. Biochimie **69**:215–221.
17. **Touati, E., F. Dassa, and P. L. Boquet.** 1986. Pleiotropic mutations in *appR* reduce pH 2.5 expression and restore succinate utilisation in CRP-deficient strains of *Escherichia coli*. Mol. Gen. Genet. **202**:257–264.
18. **Touati, E., E. Dassa, J. Dassa, and P. L. Boquet.** 1987. Acid phosphatase (pH 2.5) of *Escherichia coli*: regulatory characteristics, p. 31–40. *In* A. Torriani-Gorini, F. G. Rothman, S. Silver, A. Wright, and E. Yagil (ed.), Phosphate metabolism and cellular regulation in microorganisms. American Society for Microbiology, Washington, D.C.
19. **von Heijne, G.** 1983. Patterns of amino acids near the signal-sequence cleavage site. Eur. J. Biochem. **133**:17–21.
20. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phase cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33**:103–119.