Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 2462–2466, May 1983
Biochemistry

Nucleotide sequence of the structural gene (*pyrB*) that encodes the catalytic polypeptide of aspartate transcarbamoylase of *Escherichia coli*

(bicistronic operon/*pyrBI*/deduced amino acid sequence)

TIMOTHY A. HOOVER, WILLIAM D. ROOF, KAREN F. FOLTERMANN, GERARD A. O'DONOVAN, DAVID A. BENCINI, AND JAMES R. WILD*

Department of Biochemistry and Biophysics, Texas A & M University, College Station, Texas 77843

Communicated by William N. Lipscomb, January 14, 1983

ABSTRACT **The deoxyribonucleotide sequence of *pyrB*, the cistron encoding the catalytic subunit of aspartate transcarbamoylase (carbamoylphosphate:L-aspartate carbamoyltransferase, EC 2.1.3.2), has been determined. The *pyrB* gene encodes a polypeptide of 311 amino acid residues initiated by an NH_2-terminal methionine that is not present in the catalytically active polypeptide. The DNA sequence analysis revealed the presence of an eight-amino-acid sequence beginning at Met-219 that was not detected in previous analyses of amino acid sequence. This octapeptide sequence provides an additional component of the disordered loop in the equatorial domain of the catalytic polypeptide. It had been found previously that the catalytic polypeptide is expressed from a bicistronic operon that also produces the regulatory polypeptide encoded by *pyrI*. A single transcriptional control region precedes the structural gene of the catalytic polypeptide and a simple 15-base-pair region separates its COOH terminus from the structural gene of the regulatory polypeptide. The chain-terminating codon of the catalytic polypeptide may contribute to the ribosomal binding site for the regulatory polypeptide and thus assist coordinate expression of the two cistrons.**

The aspartate transcarbamoylase holoenzyme (ATCase; aspartate carbamoyltransferase; carbamoylphosphate:L-aspartate carbamoyltransferase, EC 2.1.3.2) is a dodecamer composed of two catalytic trimers (c_3) and three regulatory dimers (r_2) (1). This oligomer, $2(c_3)\cdot3(r_2)$, is so arranged that each catalytic monomer is in contact with three other catalytic chains and two regulatory chains, while each regulatory monomer interfaces with one other regulatory chain and two catalytic chains (2). Perbal and Hervé demonstrated that the two polypeptides are produced in approximately balanced biosyntheses (3), and we recently demonstrated that the two genes were closely linked and encoded on a 6.0-kilobase (kb) fragment isolated from $\lambda dargI^+pyrB^+$ transducing phage obtained from N. Glansdorff (4). The structural gene that encodes the catalytic polypeptide was designated *pyrB* (5) and the structural gene encoding the regulatory polypeptide has been designated *pyrI* (4, 6). Recently, we demonstrated that *pyrBI* is organized as a bicistronic operon possessing a single control region, which includes a ρ-independent attenuator sequence, a region of dyad symmetry overlapping transcriptional initiation, and a presumed leader polypeptide whose termination overlaps the ribosomal binding site for the translation of the catalytic polypeptide (7).

The architecture of ATCase from *Escherichia coli* has been extensively examined because the regulatory controls that affect the catalytic sites are mediated by distinct allosteric sites located on separate regulatory polypeptides. ATCase catalyzes the first reaction unique to pyrimidine biosynthesis, the condensation of carbamoyl phosphate and aspartate to produce carbamoyl-L-aspartate and orthophosphate (8). The enzyme is subject to allosteric inhibition by CTP, one of the ultimate products of pyrimidine ribonucleotide biosynthesis (9); in contrast, ATP is a heterotropic activator of ATCase from *E. coli* (10) and competes with CTP for the same binding sites on the regulatory subunits (11). The appropriate heterotropic effects of these nucleotide effectors occur as a consequence of conformational changes (12, 13) in the multimeric enzyme that alter the homotropic interactions of both substrates, aspartate and carbamoylphosphate (9, 10). The kinetic and physical properties of native ATCase holoenzyme and its constituent subunits from *E. coli* have been extensively examined and atomic models describing the biochemical and crystallographic structure continue to be refined (14–16). The recent work of Honzatko *et al.* (2) represents the current state of the art. Nonetheless, the efforts to describe the functional architecture of ATCase completely have been hampered by the absence of a published amino acid sequence of the catalytic polypeptide (see ref. 17). The amino acid sequence of the regulatory polypeptide was published by Weber in 1968 (18). In this paper we report the nucleotide sequence of the structural gene for the catalytic polypeptide of ATCase and the amino acid sequence deduced from the DNA sequence.

MATERIALS AND METHODS

DNA Sequence Analysis. DNA fragments were isolated from plasmid pPB-h105 (*p*lasmid *p*yrimidine-*B* cistron *h*oloenzyme strain *105*) by restriction endonuclease digestion followed by electrophoresis on polyacrylamide gels as described (7). The fragments were dephosphorylated with bacterial alkaline phosphatase and labeled at their 5′ ends with [γ-^{32}P]ATP by utilizing phage T4 polynucleotide kinase as described by Maxam and Gilbert (19). Sequence determination of appropriate overlapping fragments was accomplished by using the C, C+T, A+G, A>C, and G specific cleavages defined by Maxam and Gilbert (19) except that the A+G reactions were performed at 37°C as modified by Smith and Calvo (20). Sequencing gels were prepared as described by Maxam and Gilbert for 8% gels (19) and by Smith and Calvo for 5% gels (20).

Materials. Restriction endonucleases were obtained from Bethesda Research Laboratories, New England BioLabs, or Boehringer Mannheim and used according to the suppliers' recommendations. Bacterial alkaline phosphatase and T4 polynucleotide kinase were purchased from Bethesda Research

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: ATCase, aspartate transcarbamoylase; c and r, catalytic and regulatory polypeptides of ATCase; kb, kilobase.

* To whom reprint requests should be addressed.

Laboratories and used as described by Maxam and Gilbert (19). Biochemical reagents were purchased from New England Nuclear, Calbiochem–Behring, or Sigma.

Bacterial Strains and Plasmids. F'-plasmid DNA was purified from *E. coli* K-12 (KLF17/KL132) carrying F'117 (F' *pyrBI*$^+$ *arg*$^+$), and recombinant DNA plasmids of various construction were produced by ligation of appropriate restriction fragments into pBR322 as described (7). Plasmids were maintained in *E. coli* TB2 (K-12 Δ*argI–pyrBI*), which was derived from the insertion and subsequent excision of phage Mu dl (*lac* Ap^r) in *pyrB* of the *E. coli* K-12 chromosome according to the procedures of Casadaban and Cohen (21). Bacterial culture, plasmid isolation, and the preparation of cell-free extracts were accomplished by standard procedures (7, 22). The plasmid used in these studies is a 2.8-kb fragment of F'117 inserted into the *Sal* I and *Pst* I sites of pBR322. This plasmid carries both *pyrB* and *pyrI* with their complete control region (7) and produces ATCase holoenzyme [M_r = 310,000; $2(c_3)\cdot3(r_2)$] at approximately 20-fold higher levels than the K-12 parental strain. Enzyme assays for ATCase in cell-free extracts and estimation of molecular weights by ascending Sephadex G-200 chromatography were performed by described methods (7, 22).

RESULTS AND DISCUSSION

Nucleotide Sequence. The *pyrB* cistron is located at approximately 96.0 min on the *E. coli* chromosome (5) and has been mapped contiguous with *pyrI* and adjacent to *argI* with approximately 3,000 base pairs separating the genes (Fig. 1). The 2.8-kb fragment defined by the bracketed region bordered by *Pst* I and *Sal* I sites contains the *pyrBI* operon and controlling regions (7). The nucleotide sequence was determined by the use of base-specific chemical modifications of small restriction fragments according to the scheme presented in Fig. 1. Previously, we determined the nucleotide sequence of the 200-base-pair promoter–leader region of *pyrBI* and the adjacent NH_2-terminal portion of the catalytic polypeptide (7). Pauza *et al.* (23) have recently determined the nucleotide sequence of the COOH-terminal region of the catalytic polypeptide and the intercistronic region and the NH_2-terminal portion of the regulatory polypeptide. The total nucleotide sequence of *pyrB*, including its putative ribosomal binding site for the catalytic polypeptide, is shown in Fig. 2. The NH_2 terminus of the catalytic polypeptide was described by Gigot *et al.* in 1977 (24) as Ala-Asn-Pro-Leu-Tyr-Gln-Lys-His-Ile-Ile . . ., and this sequence corresponded with the polypeptide sequence deduced from the nucleotide sequence shown in Fig. 2. The DNA sequence predicted the translation of an NH_2-terminal methionine that was not present in the mature catalytic polypeptide and the second residue (alanine) assumed that position. The size of the polypeptide predicted from the nucleotide sequence was 311 amino acids, with an expected molecular mass of 34,074 daltons, including the NH_2-terminal methionine, or 33,943 daltons without that methionine. Those values correspond remarkably well with the estimate of 35,000 daltons obtained by physicochemical measurements of the purified polypeptide (25).

A comparison of the deduced amino acid sequence and the proposed amino acid sequence of the initial catalytic polypep-

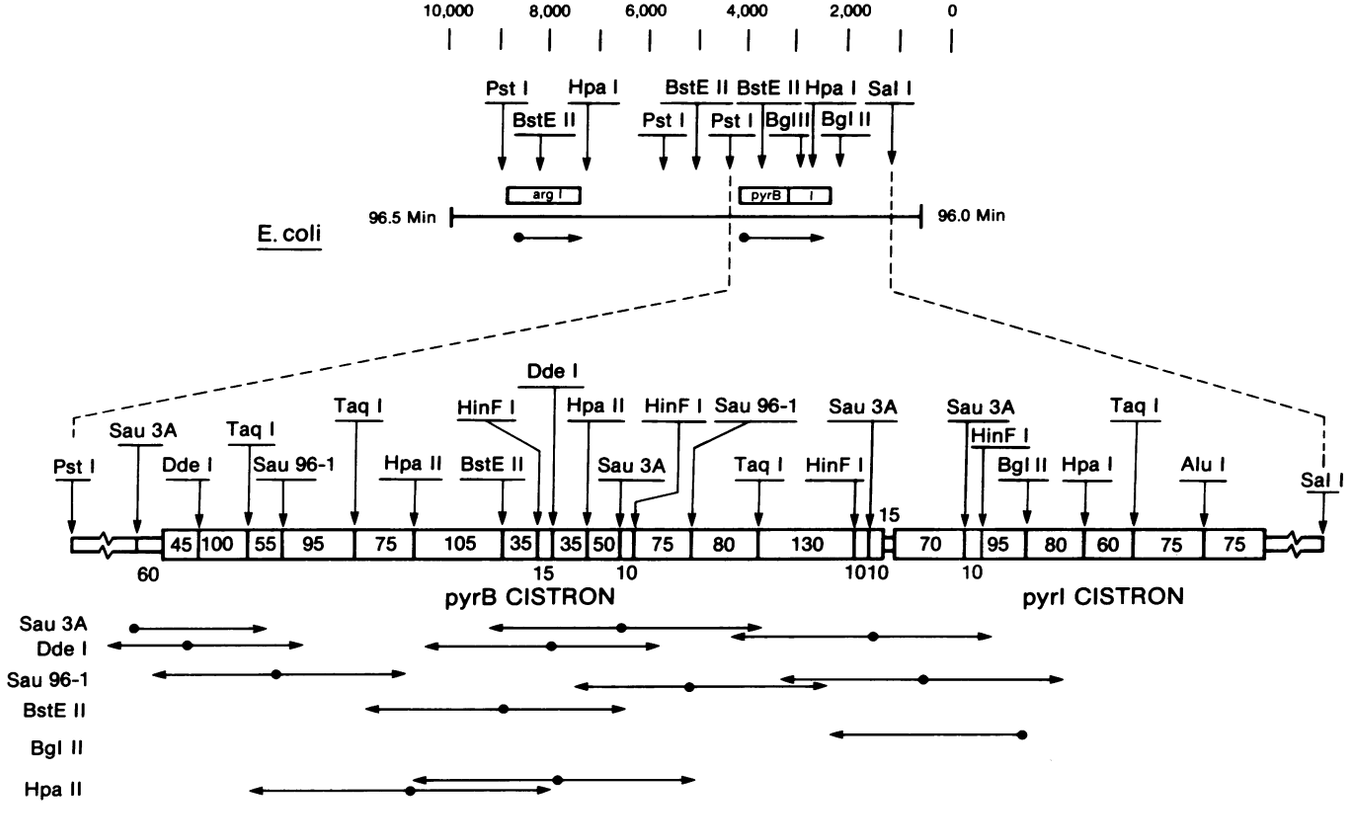

FIG. 1. Restriction map and sequencing strategy for the analysis of the *E. coli pyrB* cistron, which encodes the catalytic polypeptide of ATCase. The appropriate region of the *E. coli* chromosome is represented at the top of the figure in approximate numbers of base pairs. An expanded restriction map is presented for the surrounding chromosomal region as subcloned on a 2.8-kb restriction fragment bordered by *Pst* I and *Sal* I sites. Six different restriction endonuclease digests were used to develop the nucleotide sequence by the procedures of Maxam and Gilbert (19) as described in the text. A detailed restriction endonuclease map is defined by representing the size (in base pairs) of various restriction fragments within the horizontal bar diagram. The vertical arrows indicate the specific locations of restriction sites. The horizontal arrows beneath the bar diagram demonstrate the direction and length of various sequencing reactions.

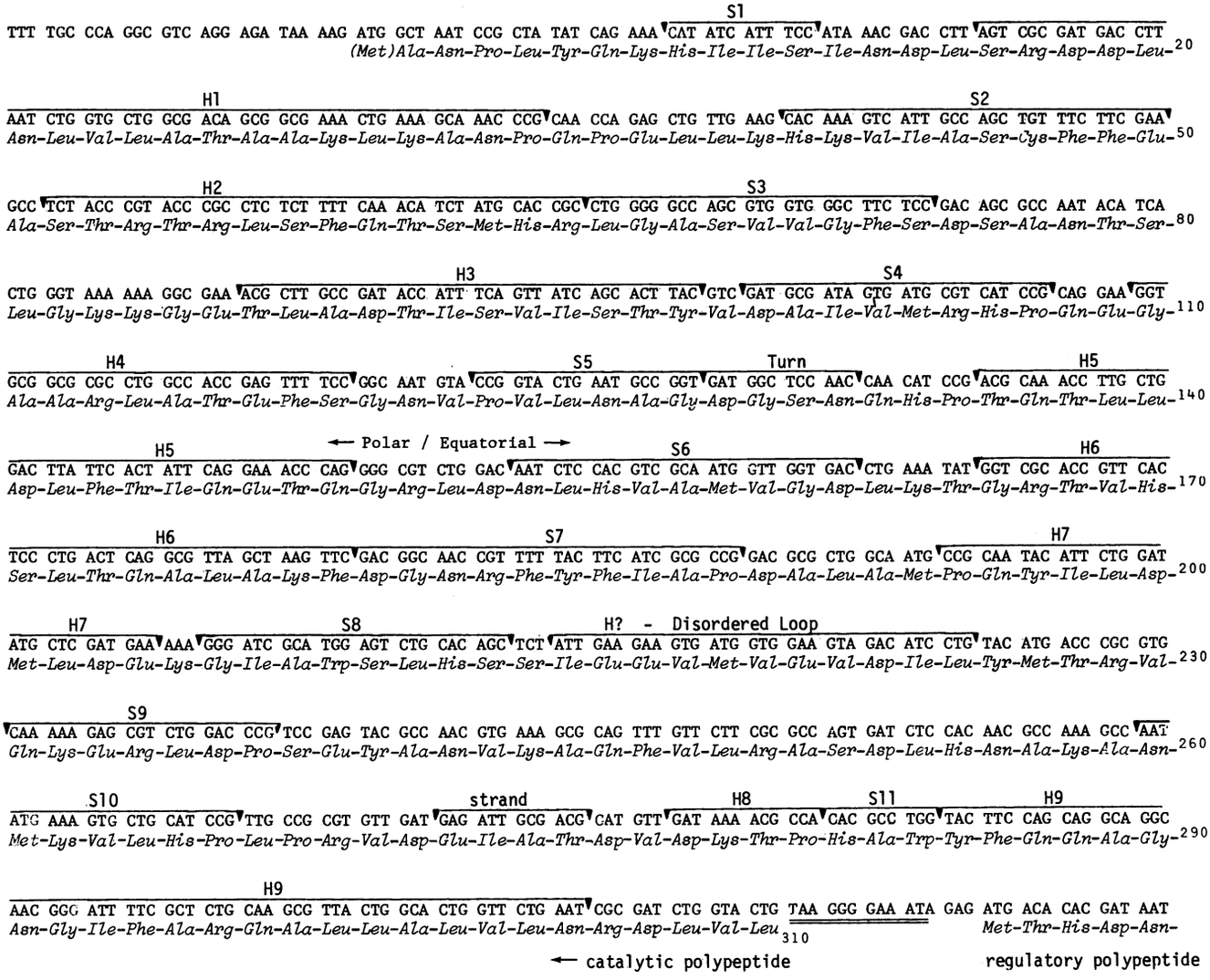

FIG. 2. Nucleotide sequence of the *pyrB* gene and flanking regions. The DNA sequence was used to deduce the amino acid sequence of the catalytic polypeptide of ATCase. The amino acid residues are numbered on the right side. The NH_2 terminus of the regulatory polypeptide is depicted on the last line, beginning with Met-Thr-His-Asp-Asn. The presumed ribosomal binding sites are indicated by solid lines drawn under the appropriate nucleotide sequences (line one and the last line represent the catalytic and regulatory sites, respectively). Brackets above the nucleotide sequence are used to identify those regions that code for various α-helix and β-sheet regions of ATCase as identified by Monaco *et al.* (16) and Honzatko *et al.* (2). The polar and equatorial domains of the enzyme are generally defined by the first and last half of the gene, respectively. (This observation is indicated at nucleotide 510, corresponding to amino acid residues Gln-149–Gly-150.)

tide of ATCase determined by Edman degradation (17) revealed the presence of an insert of eight amino acids at Met-219. The octapeptide is bounded by Met-219 and Met-227 and occupies the disordered loop of the catalytic polypeptide shown in Fig. 3 at the top of the equatorial domain. This sequence, Met-Val-Glu-Val-Asp-Ile-Leu-Tyr (Fig. 2), has a high probability of adopting an α-helical structure when analyzed according to the assumptions of Chou and Fasman (26); however, that helix does not appear in the equatorial domain (W. N. Lipscomb, personal communication). Nonetheless, this addend is located between β-sheet structures termed S8 and S9 (see Fig. 3). Table 1 presents a comparison of the potential secondary structures identified by Honzatko *et al.* (2) and the deduced sequence reported in Fig. 2. The preliminary amino acid primary sequence (17) was assigned to the type of secondary structure deduced from x-ray crystallographic analysis and the α-helical structures and the β-sheet regions ("strand," S*n*). These are identical through S8 (Gly-206 to Ser-213) until the disordered loop containing the cryptic octapeptide sequence. After that point the amino acid sequence and the requirements of the x-ray crystallographic analyses began to diverge. In order to approximate the amino acid sequence to the atomic model of ATCase based on approximation of α-carbon positions, Honzatko *et al.* (2) added three residues in numerical description after residue 242. We have reassigned the strand and helix types of secondary structure to their appropriate amino acid location deduced from DNA sequence analysis.

Various studies have suggested a larger catalytic polypeptide than previously thought. The unpublished primary sequence from Konigsberg's laboratory has been frequently cited and contained 302 amino acids. However, the analysis of the deduced amino acid sequence and the proposed secondary structure presented by Honzatko *et al.* (2) suggested that the polypeptide had to possess at least 305 amino acids. Landfear *et al.* (27) identified an 11-amino-acid sequence that seemed to correspond to the amino acid residues Leu-208 to Lys-217 derived from the sequence of Konigsberg and co-workers (reported as a personal communication); however, this sequence actually

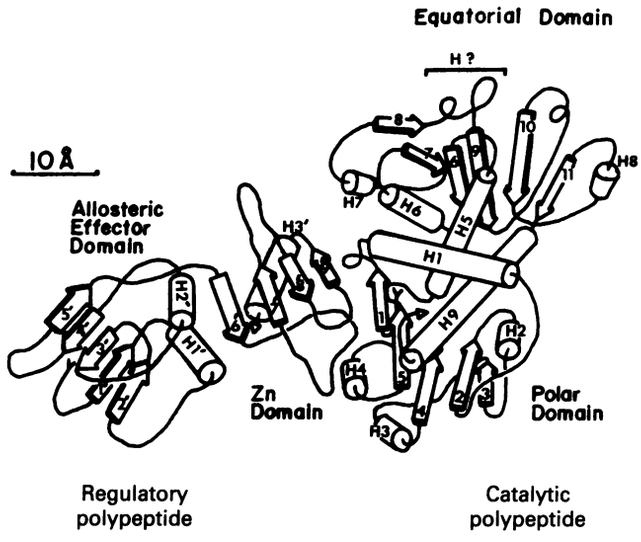

FIG. 3. Diagram of the structure of a single regulatory catalytic unit as estimated from the α-carbon positions determined by electron density analysis of x-ray crystallographic studies. This figure is modified from Honzatko *et al.* (2) with permission. The helices of the regulatory and catalytic polypeptides are represented by cylinders identified as H1'–H3' for the regulatory chain and H1–H9 for the catalytic chain. The octapeptide whose presence was confirmed in these studies is indicated as H? in the equatorial domain. Similarly, β-sheet regions are represented as 1'–9' or 1–11. The various functional domains of the enzyme are indicated. The potential helix formation for H? is indicated beneath the diagram (26).

Table 1. Comparison of secondary structure determined by amino acid sequence and by nucleotide sequence

Element name*	Type of secondary structure*	Amino acid primary sequence†	Location from nucleotide sequence‡
S1	Strand	His-8–Ser-11	His-8–Ser-11
H1	Helix	Ser-16–Pro-34	Ser-16–Pro-34
S2	Strand	His-41–Glu-50	His-41–Glu-50
H2	Helix	Ser-52–Arg-65	Ser-52–Arg-65
S3	Strand	Leu-66–Ser-74	Leu-66–Ser-74
H3	Helix	Thr-87–Tyr-98	Thr-87–Tyr-98
S4	Strand	Asp-100–Pro-107	Asp-100–Pro-107
H4	Helix	Gly-110–Ser-119	Gly-110–Ser-119
S5	Strand	Pro-123–Gly-128	Pro-123–Gly-128
—	Turn	Asp-129–Asn-132	Asp-129–Asn-132
H5	Helix	Thr-136–Glu-149	Thr-136–Gln-149
S6	Strand	Asn-154–Asp-162	Asn-154–Asp-162
H6	Helix	Gly-166–Phe-179	Gly-166–Phe-179
S7	Strand	Asp-180–Pro-189	Asp-180–Pro-189
H7	Helix	Pro-195–Glu-204	Pro-195–Glu-204
S8	Strand	Gly-206–Ser-213	Gly-206–Ser-213
—	Helix	Ile-216–Arg-221	Ile-214–Leu-224
S9	Strand	Gln-223–Pro-229	Gln-231–Pro-237
—	Strand	AA-244–AA-248	—— ——
S10	Strand	Asn-255–Pro-261	Asn-260–Pro-266
—	Strand	Glu-267–Thr-270	Glu-272–Thr-275
H8	Helix	Asp-271–Pro-275	Asp-278–Pro-281
S11	Strand	His-277–Trp-279	His-282–Trp-284
H9	Helix	Tyr-280–Asn-300	Tyr-285–Asn-305

* From Honzatko *et al.* (2). "Strand" indicates β-sheet and "Helix" represents α-helix (26).
† From the revised sequence of Konigsberg up to residue 242 as modified by Honzatko *et al.* (2).
‡ See Fig. 2.

corresponded to Leu-235 to Lys-244 (Fig. 2). In addition, Wall *et al.* (28) identified the amino acid sequence of a 38-residue peptide that did correspond with Leu-114 to Arg-151. These observations are all consistent with the deduced amino acid sequence reported in this paper. Konigsberg and Henderson have reevaluated their amino acid sequence data by using HPLC separation and they have supported the deduced sequence (17). The mature, catalytically active polypeptide of ATCase is composed of 310 amino acids, with alanine serving as the NH_2 terminus and a potentially free-moving five-residue sequence (Arg-Asp-Leu-Val-Leu) following helix 9 (H9; see Fig. 3) providing the COOH terminus.

The actual DNA sequence of the COOH-terminal portion of the catalytic cistron, intercistronic region, and NH_2-terminal portion of the regulatory cistron verified the reported partial sequence of the region (23) and corresponded with the published amino acid sequence of the regulatory polypeptide (18). The nucleotide sequence of the COOH-terminal region (Asn-260 to Gly-290) has been determined on only one strand.

Amino Acid Composition and Codon Usage. We have determined the amino acid composition of the catalytic polypeptide and the distribution of residues by codon usage (data not shown). The DNA sequence of the structural gene region of *pyrB* is approximately 48% A+T, which is similar with the base composition of the host *E. coli* K-12 (31). The pattern of codon usage is comparable to that observed for a mixture of several *E. coli* proteins as described by Greene *et al.* (30). Post and Nomura have suggested that the ribosomal proteins utilize the major tRNA species in order to allow the most efficient translation of these abundant proteins (29). Similarly, it appears that the catalytic cistron of ATCase utilized the most abundant species of tRNA, with the possible exception of AUA for isoleucine (3/16 isoleucine codons represent minor species).

Ribosomal Binding Sites. Although the *pyrB* and *pyrI* gene products are apparently expressed from a single mRNA, that does not ensure their equivalent translation in order to provide stoichiometric production of catalytic and regulatory polypeptides. Imamoto and Yanofsky (32) observed that the five gene products of the *trp* operon of *E. coli* are normally synthesized in equimolar amounts from its polycistronic message. However, Selker and Yanofsky (33) proposed that the observed overlapping of stop and start codons of sequentially expressed cistrons in the *trp* operon may provide a means for the translational coupling of expression. Furthermore, this might reflect a general mechanism for the production of polypeptides required in equivalent amounts (34). The translational stop (UAA) of the catalytic polypeptide was observed to overlap the ribosomal binding site for the regulatory polypeptide and provides an analogous mechanism to *trp* expression, in which the translational expression of the two cistrons may be clearly coupled (23).

Conclusions. The complete amino acid sequence of the catalytic polypeptide of ATCase has been derived from the DNA sequence of a cloned *pyrB* gene from *E. coli*. The sequence has been confirmed by comparison with previously published segments of the polypeptide (24, 27, 28) and the amino acid sequence generated by Konigsberg and Henderson (17). Our initial analysis revealed an undetected eight-amino-acid sequence corresponding to residues Met-219 to Tyr-226. This sequence appears to participate in a disordered loop assigned by Monaco *et al.* (16) and Honzatko *et al.* (2) to the equatorial domain as determined by x-ray structure analysis. The sequence allows

the clarification of the regions involved in secondary structure formation by Lipscomb and co-workers (11–13, 16).

The codons used for various amino acids are translated by major species of isoaccepting tRNAs (29, 30). Thus it would seem that the catalytic polypeptide would be rapidly translated if its ribosomal binding site was accessible. Furthermore, the catalytic polypeptide chain-terminating codon may participate in the ribosomal binding site of the regulatory polypeptide. These interactions might provide some modulation for the apparently coordinate expression of *pyrB* and *pyrI* (3). The nature of the regulatory controls affecting *pyrBI* expression invites further analysis.

This research has been supported by the Robert A. Welch Foundation (A-915), the National Institutes of Health (NIHMS-5R01GM29152), the National Science Foundation (PCM-8021983), and the Texas Agricultural Experiment Station.

1. Cohlberg, J. A., Pigiet, V. P., Jr., & Schachman, H. K. (1972) *Biochemistry* **11**, 3396–3411.
2. Honzatko, R. B., Crawford, J. L., Monaco, H. L., Ladner, J. E., Edwards, B. F. P., Evans, D. R., Warren, S. G., Wiley, D. C., Ladner, R. C. & Lipscomb, W. N. (1982) *J. Mol. Biol.* **160**, 219–263.
3. Perbal, B. & Hervé, G. (1972) *J. Mol. Biol.* **70**, 511–529.
4. Wild, J. R., Foltermann, K. F., Roof, W. D. & O'Donovan, G. A. (1981) *Nature (London)* **292**, 373–375.
5. Bachmann, B. J. & Low, K. B. (1980) *Microbiol. Rev.* **44**, 1–56.
6. Feller, A., Piérard, A., Glansdorff, N., Charlier, D. & Crabeel, M. (1981) *Nature (London)* **292**, 370–373.
7. Roof, W. D., Foltermann, K. F. & Wild, J. R. (1982) *Mol. Gen. Genet.* **187**, 391–400.
8. Yates, R. A & Pardee, A. B. (1956) *J. Biol. Chem.* **221**, 743–756.
9. Gerhart, J. C. & Pardee, A. B. (1962) *J. Biol. Chem.* **237**, 891–896.
10. Bethell, M. R., Smith, K. E., White, J. S. & Jones, M. E. (1968) *Proc. Natl. Acad. Sci. USA* **60**, 1442–1449.
11. Honzatko, R. B., Monaco, H. L. & Lipscomb, W. N. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5105–5109.
12. Kantrowitz, E. R., Pastra-Landis, S. C. & Lipscomb, W. N. (1980) *Trends Biochem. Sci.* **5**, 124–128.
13. Kantrowitz, E. R., Pastra-Landis, S. C. & Lipscomb, W. N. (1980) *Trends Biochem. Sci.* **5**, 150–153.
14. Gerhart, J. C. (1970) *Curr. Top. Cell. Regul.* **2**, 275–325.
15. Jacobson, G. R. & Stark, G. R. (1975) *J. Biol. Chem.* **250**, 6852–6860.
16. Monaco, H. L., Crawford, J. L. & Lipscomb, W. N. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 5276–5280.
17. Konigsberg, W. H. & Henderson, L. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2467–2471.
18. Weber, K. (1968) *J. Biol. Chem.* **243**, 543–546.
19. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
20. Smith, P. R. & Calvo, J. M. (1980) *Nucleic Acids Res.* **9**, 3365–3377.
21. Casadaban, M. J. & Cohen, S. N. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4530–4533.
22. Wild, J. R., Foltermann, K. F. & O'Donovan, G. A. (1980) *Arch. Biochem. Biophys.* **201**, 506–517.
23. Pauza, C. D., Karels, M. J., Navre, M. & Schachman, H. K. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 4020–4024.
24. Gigot, D., Glansdorff, N., Legrain, C., Piérard, A., Stalon, V., Konigsberg, W., Caplier, I., Strosberg, A. D. & Hervé, G. (1977) *FEBS Lett.* **81**, 28–32.
25. Meighen, E. A., Pigiet, V. & Schachman, H. K. (1970) *Proc. Natl. Acad. Sci. USA* **65**, 234–241.
26. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol.* **47**, 45–148.
27. Landfear, S. M., Evans, D. R. & Lipscomb, W. N. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 2654–2656.
28. Wall, K. A. & Schachman, H. K. (1979) *J. Biol. Chem.* **254**, 11917–11926.
29. Post, L. E. & Nomura, M. (1980) *J. Biol. Chem.* **255**, 4660–4666.
30. Greene, P. J., Gupta, M., Boyer, H. W., Brown, W. E. & Rosenberg, J. M. (1981) *J. Biol. Chem.* **256**, 2143–2153.
31. Normore, W. M. & Brown, J. R. (1970) *CRC Handbook of Biochemistry, Selected Data For Molecular Biology*, ed. Sober, H. A. (CRC, Cleveland, OH), 2nd Ed., pp. H24–H74.
32. Imamoto, F. & Yanofsky, C. (1967) *J. Mol. Biol.* **28**, 25–35.
33. Selker, E. & Yanofsky, C. (1979) *J. Mol. Biol.* **130**, 135–143.
34. Oppenheim, D. S. & Yanofsky, C. (1980) *Genetics* **95**, 785–795.