Proc. Natl. Acad. Sci. USA
Vol. 74, No. 4, pp. 1507–1510, April 1977
Biochemistry

The amino acid sequence of β-galactosidase of *Escherichia coli*

(*lac* operon/protein sequencing)

AUDREE V. FOWLER AND IRVING ZABIN

Department of Biological Chemistry, University of California at Los Angeles School of Medicine, and Molecular Biology Institute, University of California, Los Angeles, California 90024

Communicated by Emil L. Smith, January 31, 1977

ABSTRACT **The amino acid sequence of β-galactosidase was determined. The protein contains 1021 amino acid residues in a single polypeptide chain. The subunit molecular weight calculated from the sequence is 116,248. The sequence determination, carried out mainly by conventional methods, was aided by complementation tests, by the use of termination mutant strains, and by a new immunochemical method. The five residue sequence Thr-Pro-His-Pro-Ala appears twice within the polypeptide chain, but no other striking homologous features are evident.**

β-Galactosidase (β-D-galactoside galactohydrolase, EC 3.2.1.23) is specified by the first structural gene (*lac* Z) of the *lac* operon in *Escherichia coli*. Physical and chemical studies have shown that the protein is a tetramer of four identical, unusually long, polypeptide chains. Estimates of the size of the monomer have varied from about 1000 to 1200 amino acid residues; the value of 1170 has been assumed in the past (1, 2).

Although the determination of the primary structure of β-galactosidase was a major undertaking, it seemed warranted for a number of reasons. Sequence information is important in order to correlate some of the extensive genetic data available on the Z gene with the protein, to investigate enzyme structure-function relationships, and to study the origin of this single protein and other proteins of the *lac* operon by examination for homology.

The amino acid sequence of β-galactosidase has now been completed and is presented here.

RESULTS AND DISCUSSION

The amino acid sequence proposed for β-galactosidase is shown in Fig. 1. From the composition (Table 1) molecular weights of 116,248 for the monomer and 464,992 for the tetramer were calculated.

The sequence was derived by studies of peptides obtained by cleavage of the protein with trypsin, chymotrypsin, and cyanogen bromide (CNBr). Structure determination was initiated by isolation of tryptic peptides (3, 4) including the amino- and carboxyl-terminal fragments (5). Additional large peptides were obtained from a tryptic digest of β-galactosidase blocked at lysine residues with citraconic anhydride. Details of peptide isolation and sequence determination will be published elsewhere.

Of the 24 unique peptides produced by cyanogen bromide treatment, 8 ranging in size from 2 to 15 residues were purified by standard techniques of paper electrophoresis and paper chromatography. The 16 larger peptides, containing 23 to 119 residues, were chromatographed at pH 5.0 on a *O*-carboxymethylcellulose column in 0.02 M ammonium acetate buffer containing 8 M urea and were eluted with a salt gradient (6). The elution position of these peptides can be seen in Fig. 2. Some of the peaks in the profile represent fragments obtained in low yield which were not cleaved at certain methionine residues, or peptides derived by cleavage of the three aspartyl-prolyl bonds in β-galactosidase. All peptides were purified further by gel filtration and, in some cases, by additional ion-exchange chromatography procedures (6). Criteria of purity included dansyl amino-terminal analysis, electrophoresis on 7.5% polyacrylamide gels containing urea, and automated sequence analysis.

The structure of small peptides was obtained by manual methods. The larger peptides were analyzed in a Beckman Sequenator by using the 0.1 M Quadrol program with dual benzene/ethyl acetate wash with some modifications (ref. 7 and Beckman program 030176, courtesy of Jack Ohms). Excellent results were obtained in most cases. For example, 52 residues of the 61 in CNBr21 were identified. All CNBr peptides were also cleaved with trypsin. In some cases, additional cleavages with chymotrypsin, thermolysin and/or staphylococcal protease were necessary to establish the complete sequences of the CNBr peptides. Carboxypeptidase A was used to establish carboxyl-terminal sequences.

Cyanogen bromide peptides were placed in order by comparison to sequences in tryptic and chymotryptic peptides as indicated in Fig. 1. The order CNBr5-CNBr6 was confirmed by isolation of a chymotryptic peptide containing residues 204–209. CNBr13–CNBr14 are the only peptides joined by a one residue overlap.

Sequence order determination was also aided by other techniques, such as α-complementation. When a CNBr digest is added to an extract of the genetically-defined deletion mutant strain M15, which produces a defective β-galactosidase, enzyme activity is restored (8). The purification of a single peptide CNBr2, residues 3–92, was monitored for activity in this manner (9).

Another aid for determining the order of the peptides was the use of termination mutants. The polypeptide from strain NG125 that maps near the center of the *lac* Z gene has a molecular weight of approximately 60,000 (10, 11). A cyanogen bromide digest of this polypeptide was chromatographed on a *O*-carboxymethylcellulose column using conditions identical to those used for a digest of the whole protein. The elution profile was thus a kind of fingerprint, and peptides which were identified by automated sequence analysis could be assigned to the amino-terminal half of the molecule.

A new immunochemical method was also devised. Antibodies were prepared against many cyanogen bromide peptides and were used to search for overlapping peptides. For example, the binding of ^{125}I-labeled CNBr21 to antibody against CNBr21 was found to be inhibited by a tryptic digest of citraconyl β-galactosidase. Purification of an inhibiting peptide containing the carboxyl-terminal 31 residues of CNBr21 and the amino-terminal 13 residues of CNBr22 was assayed by measuring inhibition (12). This procedure saved considerable time by

Abbreviation: CNBr, cyanogen bromide.

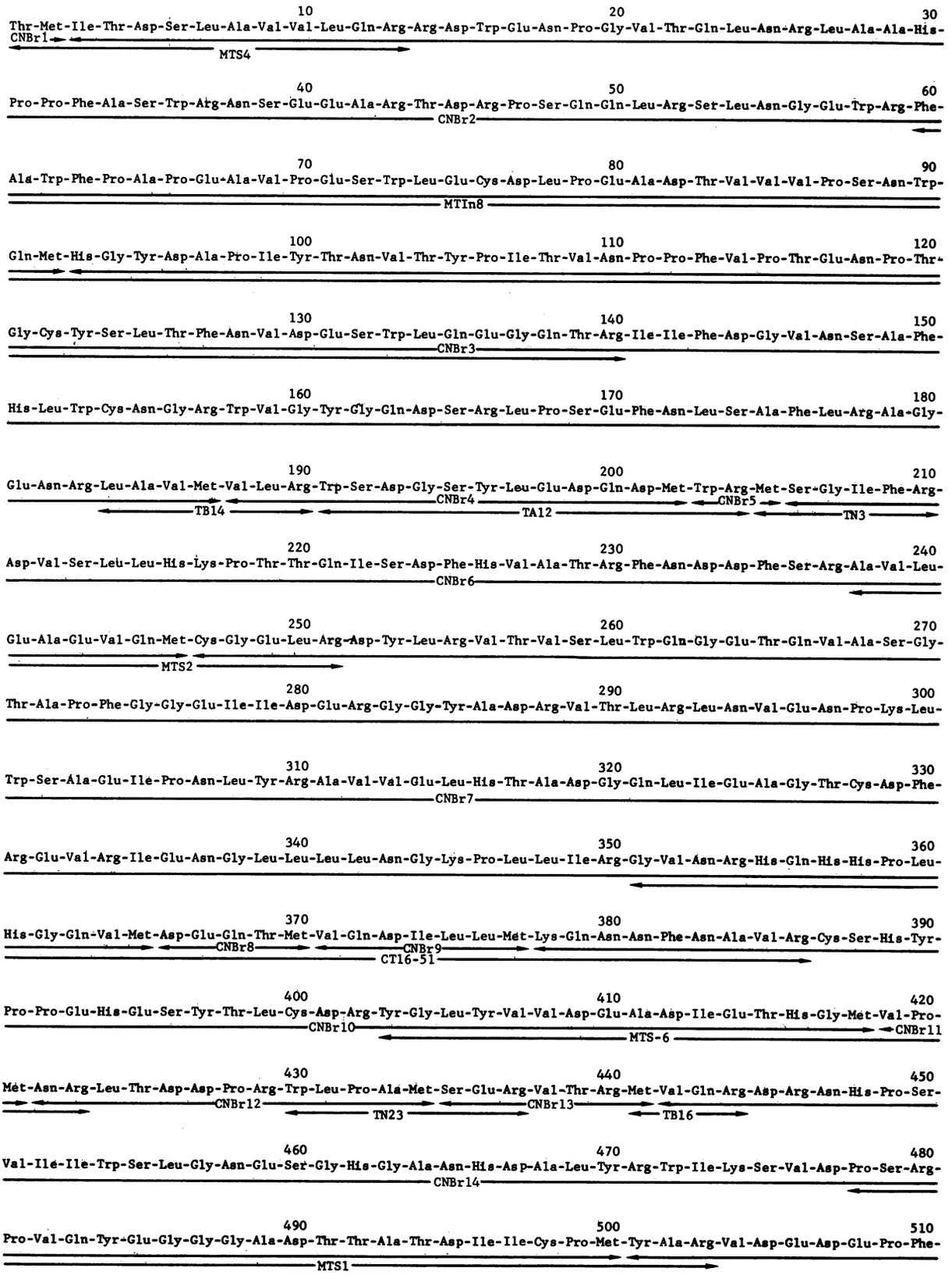

FIG. 1. *Continued on following page.*

avoiding the necessity for examination of many fractions in order to find the desired peptide.

Finally, we were aided in the sequence determination by a correlative study of the DNA sequence of the early part of the *lac* Z gene (A. Maxam and W. Gilbert, personal communication). Assignments of amino acid residues 1–145 were found to agree with the assignments predicted from the DNA sequence. Several minor uncertainties could be resolved, as for example an amide assignment at residue 135.

Completion of the sequence determination proves that there

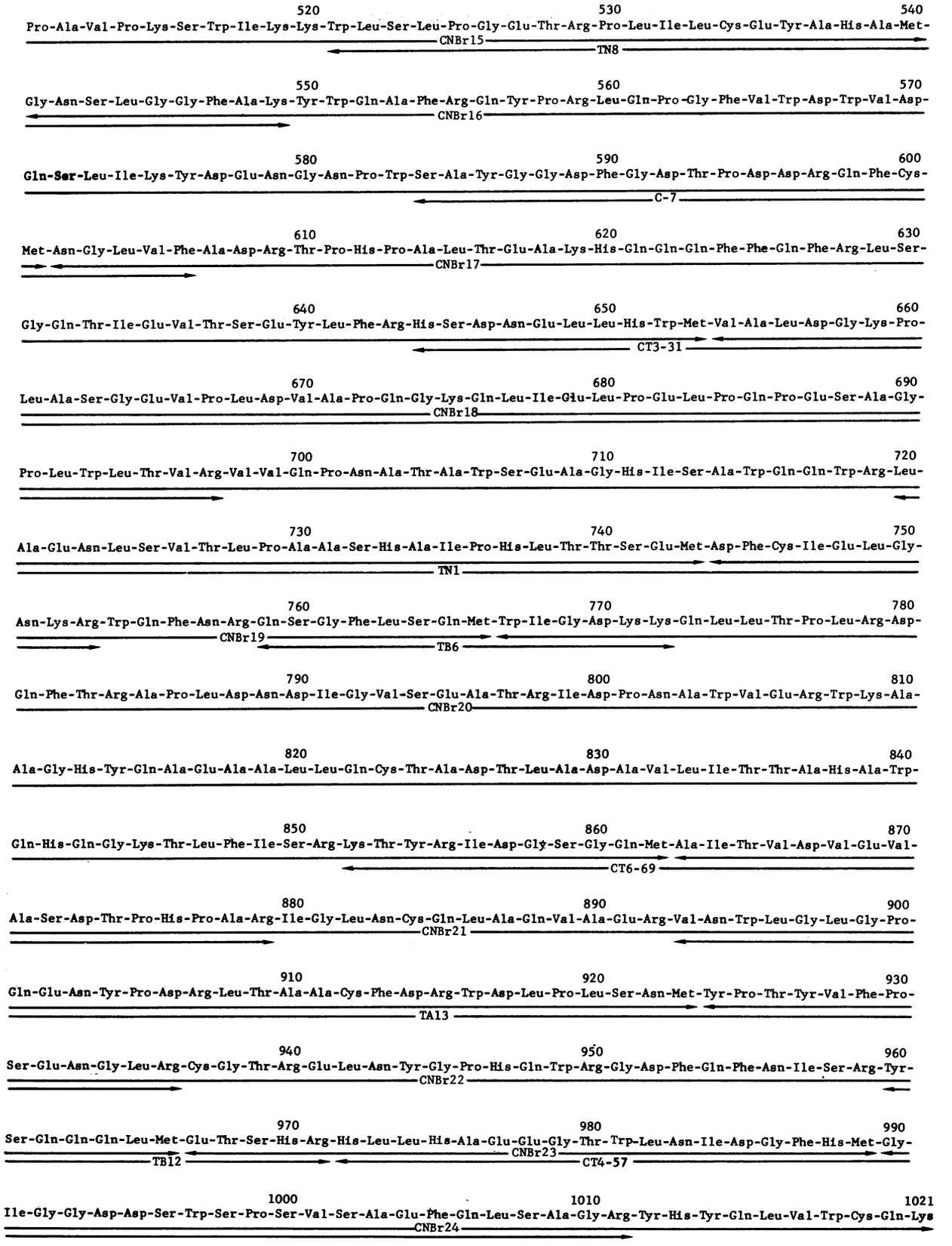

FIG. 1. Amino acid sequence of β-galactosidase. The letters CNBr indicate cyanogen bromide peptides; MTS, MTIn, TA, TB, TN, and CT refer to tryptic peptides, and C refers to chymotryptic peptides.

are no smaller subunits making up the monomer of β-galactosidase. The single polypeptide chain of 1021 residues is the largest whose primary structure has been established so far. It has several unique features besides its large size. The tryptophan content is extremely high (38 residues). The lysine content, on the other hand, is quite low (20 residues). Only five of these are in the first half of the chain, whereas eleven are between residues 515 and 772. The only lysine beyond residue 852 is at the

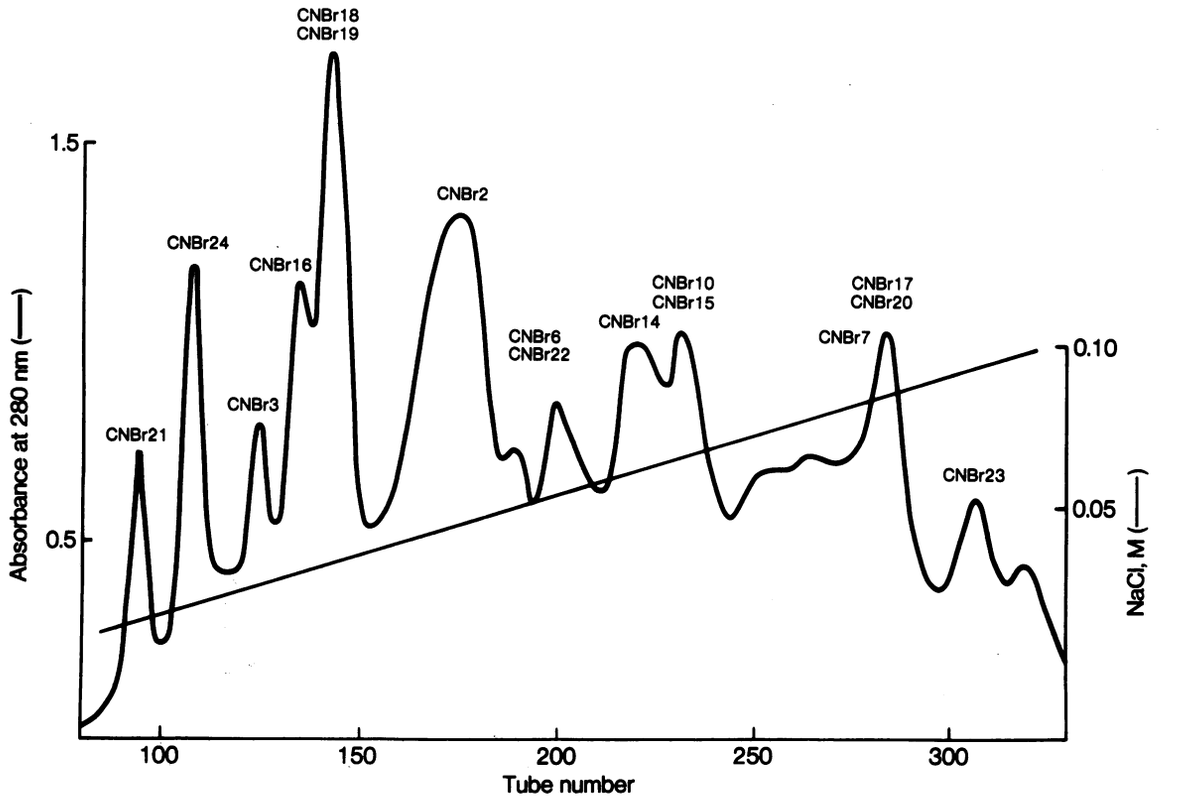

FIG. 2. *O*-Carboxymethylcellulose chromatography of cyanogen bromide peptides of β-galactosidase. Peptides were applied to a column (2.5 × 38 cm) in 0.02 M ammonium acetate, pH 5.0, and 8 M urea, and were eluted with a linear gradient of 0–0.15 M NaCl in the same buffer. Total volume was 3000 ml.

carboxyl-terminus. Because it is generally believed that lysine residues are on the exterior of protein molecules, this suggests that much of the amino- and carboxyl-terminal regions of the polypeptide chain of β-galactosidase may be buried within the molecule.

The only striking duplication within the protein is the five-residue sequence Thr-Pro-His-Pro-Ala which is present at residues 610–614 and again at residues 874–878. No other internal homologous features are obvious nor does there appear to be any significant homology of β-galactosidase with the *lac* repressor protein (J. M. Hood, A. V. Fowler, and I. Zabin, unpublished).

Table 1. Amino acid composition of β-galactosidase

Amino acid	No. residue found	
	Analysis	Sequence
Tryptophan	27	38
Lysine	23	20
Histidine	31	34
Arginine	64	66
Aspartic acid	105	110
Threonine	59	56
Serine	60	61
Glutamic acid	124	121
Proline	62	64
Glycine	72	70
Alanine	81	76
Half-cystine	15	16
Valine	64	63
Methionine	23	23
Isoleucine	38	39
Leucine	96	95
Tyrosine	29	31
Phenylalanine	38	38
Total residues	1011	1021

We thank A. J. Brake for determining the sequence of the carboxyl-terminal cyanogen bromide peptide and for help in the peptide isolation steps, and Steve Barta, Paulette Osborne, Karin Pratt, Carol Sander, and Val Schaeffer for excellent technical assistance. This investigation was supported in part by U.S. Public Health Service Grant AI 04181 and National Science Foundation Grant GB 37559.

1. Zabin, I. & Fowler, A. V. (1970) in *The Lactose Operon*, eds. Zipser, D. & Beckwith, J. R. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 27–47.
2. Wallenfels, K. & Weil, R. (1972) in *The Enzymes*, ed. Boyer, P. D. (Academic Press, New York), pp. 617–663.
3. Fowler, A. V. & Zabin, I. (1970) *J. Biol. Chem.* **245**, 5032–5041.
4. Fowler, A. V. (1972) *J. Biol. Chem.* **247**, 5425–5431.
5. Zabin, I. & Fowler, A. V. (1972) *J. Biol. Chem.* **247**, 5432–5435.
6. Fowler, A. V. (1975) in *Solid Phase Methods in Protein Sequence Analysis*, ed. Laursen, R. A. (Pierce Cemical Co., Rockford, IL), pp. 169–177.
7. Brauer, A. W., Margolies, M. N., & Haber, E. (1975) *Biochemistry* **14**, 3029–3035.
8. Lin, S., Villarejo, M., & Zabin, I. (1970) *Biochem. Biophys. Res. Commun.* **40**, 249–254.
9. Langley, K. E., Fowler, A. V. & Zabin, I. (1975) *J. Biol. Chem.* **250**, 2587–2592.
10. Fowler, A. V. & Zabin, I. (1966) *Science* **154**, 1027–1029.
11. Villarejo, M. R. & Zabin, I. (1974) *J. Bacteriol.* **120**, 466–474.
12. Brake, A. J., Celada, F., Fowler, A. V. & Zabin, I. (1977) *Anal. Biochem.* in press.