*Proc. Natl. Acad. Sci. USA*
Vol. 77, No. 11, pp. 6531–6535, November 1980
Biochemistry

# Mass spectra of partial protein hydrolysates as a multiple phase check for long polypeptides deduced from DNA sequences: $NH_2$-terminal segment of alanine tRNA synthetase

(gas chromatographic mass spectrometry/protein sequencing/DNA sequencing)

WALTER C. HERLIHY*, NANCY J. ROYAL*, K. BIEMANN*†, SCOTT D. PUTNEY‡, AND PAUL R. SCHIMMEL†‡

Departments of *Chemistry and ‡Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

*Communicated by John M. Buchanan, August 11, 1980*

**ABSTRACT** **A strategy has been developed for rapid and accurate determination of the amino acid sequence of large proteins, such as many of the members of the class of proteins known as aminoacyl tRNA synthetases. This strategy involves combining DNA sequencing of the gene for the protein of interest with gas chromatographic mass spectrometric identification of tetra- and pentapeptides in partial hydrolysates of the entire protein or very large fragments thereof. These peptides are matched to blocks of codons at locations scattered throughout the entire structural gene. Tetra- and pentapeptide sequences are sufficiently long that they are unlikely to be repeated in the protein sequence or to occur in an incorrect reading frame; therefore, they can be placed at unique clusters of codons on the DNA. This procedure rigorously establishes the proper phasing of the DNA throughout the entire length of the structural gene, and the protein sequence is thereby accurately read from the DNA sequence. This approach is being used to determine the amino acid sequence of *Escherichia coli* alanine tRNA synthetase, a protein that has approximately 900 amino acids. This paper reports the sequence of the first 165 amino acids from the $NH_2$ terminus.**

The aminoacyl tRNA synthetases are among the most intensively investigated group of enzymes from the standpoints of mechanism and of recognition of transfer RNAs (1). This is because they play a major role in protein synthesis and, in essence, by matching amino acids to anticodon trinucleotides contained within the tRNA molecules, establish the rules of the genetic code. But despite the importance of these enzymes and the extensive investigations of their mechanism of action, there is little structural information available about them. This is due in significant part to the large size of some of these enzymes. For example, many are made of single chains or subunits that are ≈1000 amino acids long (1). The sequencing of one, let alone several, enzyme(s) of this size presents a formidable task, and yet understanding many of the basic questions concerning these proteins relies heavily on such information.

The ability to sequence long DNA chains rapidly is an alternative to the conventional, but tedious and time-consuming, method of protein sequencing. The credibility of polypeptide sequences derived from DNA sequences is greatly enhanced when the $NH_2$-terminal sequence and, in addition, the COOH-terminal sequence of the protein are independently established. Nevertheless, there is considerable reluctance to accept polypeptide sequences derived from DNA sequences when the polypeptide chain is several hundred or more amino acids long and the only corroborative evidence stems from $NH_2$- and COOH-terminal sequences. Furthermore, it is often difficult to determine the amino acid sequences of polypeptide chain termini; the $NH_2$ terminus may be blocked, thus preventing standard Edman degradation procedures, and methods for sequencing the COOH terminus are inherently less useful and reliable.

Despite the accuracy of DNA sequencing and the ability to check the data obtained by sequencing both strands of the DNA, the problem remains that a single deletion or insertion error in a block of 1000 nucleotides will affect the proper reading of not one but all of the subsequent codons. This problem cannot necessarily be overcome by relying on the possibility of encountering premature stop codons in an improper reading frame. For example, in our work with alanine tRNA synthetase (AlaRS), we have found a stretch of more than 120 nucleotides within the structural gene that is devoid of stop codons in all three reading phases. Thus, there is no simple way to overcome the possibility that a deletion or insertion of a base in the DNA sequence will lead to the derivation of an incorrect amino acid sequence. Obviously, the problem becomes more serious the greater the length of the DNA.

Because of these difficulties, we have developed a strategy that ensures accurate translation of long DNA sequences over their entire lengths. Tetra- and pentapeptide sequences are determined by gas chromatographic mass spectrometric (GCMS) analysis of digests of the whole protein or very large fragments thereof. No separation of the oligopeptides is required before the analysis. In a single experiment, numerous peptide sequences are obtained from scattered regions of the protein. Because tetra- and pentapeptide sequences are unlikely to be repeated, they can be fit to unique locations in the DNA sequence. In this way, we can determine the proper phase of the DNA sequence throughout its entire length and rigorously establish the amino acid sequence.

In this paper, we describe this approach and report our initial results on the sequence of AlaRS, a protein that is ≈900 amino acids long. The primary structure of an $NH_2$-terminal segment of 165 amino acids has been elucidated by these methods.

## MATERIALS AND METHODS

**Purification and Limited Tryptic Cleavage of AlaRS; Isolation of Fragment T-1.** The source of AlaRS was a cell strain containing a hybrid plasmid (pSP101) that contains the gene for it (2). Purified enzyme was obtained by using the method of Putney *et al.* (3).

AlaRS (≈100 nmol) was dissolved in 16.0 ml of $NH_4OAc$ buffer (90 mM in OAc, pH 8.3). Trypsin was added at an enzyme-substrate ratio of 1:100 (wt/wt), and the mixture was incubated at 37°C for 3 hr. A fragment (T-1) ($M_r$ ≈40,000) was isolated from the tryptic digest by chromatography on a 2.7 ×

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: AlaRS, alanine tRNA synthetase; GCMS, gas chromatographic mass spectrometry.

† To either of whom reprint requests should be addressed.

85 cm column of Bio-Gel P-60 (Bio-Rad) eluted with 50 mM $NH_4OAc$. The T-1 pool was located by its absorbance at 280 nm and concentrated to ≈10 ml in a stirred filtration cell with a PM-10 membrane (Amicon).

**Partial Hydrolysis of T-1.** The concentrated T-1 pool (50–100 nmol) was diluted to 50 ml with either distilled water (for 6 M HCl or pepsin hydrolysis) or $NH_4OAc$ (90 mM in AcO, pH 8.3) (for all other enzymes) and reconcentrated to ≈5 ml. The T-1 solution was lyophilized before 6 M HCl hydrolysis at 110°C for 30 min. For pepsin digestion, the pH of the solution was adjusted to 3.0 with formic acid, and the solution was incubated with pepsin at an enzyme-substrate ratio of 1:50 (wt/wt) at 37°C for 1.5 hr. The digest was lyophilized, redissolved in 4 ml of $NH_4OAc$ buffer (pH 8.3), and digested with elastase for 1 hr at an enzyme-substrate ratio of 1:50 (wt/wt). Partial hydrolyses with trypsin/chymotrypsin, thermolysin, or proteinase K were performed on 50–100 nmol of T-1 in $NH_4OAc$ buffer (pH 8.3) at enzyme-substrate ratios of 1:50 to 1:100 (wt/wt). The thermolysin and proteinase K hydrolyses were incubated at 55°C for 3 hr and 30 min, respectively, and the trypsin/chymotrypsin digestion was incubated at 37°C for 14 hr. To maximize the yields of tetra- and pentapeptides, digestion conditions were first established by trial experiments monitored by $NaDodSO_4$/polyacrylamide gel electrophoresis (4). Digestions were judged to be complete when the intensity of the T-1 band was reduced to <5% of its original value.

**Conversion of Peptide Mixtures to Polyamino Alcohols; Mass Spectrometry.** After partial hydrolysis, the peptide mixtures were lyophilized and converted to the corresponding *O*-trimethylsilyl polyamino alcohols as described (5, 6). Briefly, the peptides were treated with 3 N HCl/MeOH and then with trifluoroacetic anhydride/trifluoroacetic acid, and these products were reduced with $B_2^2H_6$ (6). The reduction products were treated with trimethylsilyl diethylamine, and the mixture of *O*-trimethylsilyl derivatives was injected directly into the gas chromatograph as described (5, 6). A computer program was used for the interpretation of the mass spectral data (7).

**DNA Sequencing.** The source of DNA was a recombinant plasmid (pSP201) containing a large segment of pBR322 and of *alaS* (2). The plasmid was replicated in cell strain KL380 and isolated as described by Putney *et al.* (2). Uniquely end-labeled (with $^{32}P$) restriction fragments were obtained either by restricting doubly labeled fragments or by strand separation, and the DNA was subsequently sequenced by using the method of Maxam and Gilbert (8).

## RESULTS AND DISCUSSION

**Feasibility of Analysis of Large Polypeptides by GCMS.** To correctly translate a DNA sequence into the corresponding protein sequence it is desirable to obtain partial sequence information from throughout the protein. Ideally, this information should be obtained from the intact protein to avoid the difficult and time-consuming task of purifying a large number of peptides from a long protein chain. The GCMS peptide sequencing technique (9–11) is, with certain modifications, well suited for this purpose. It involves the partial hydrolysis of a polypeptide having 50 or fewer residues to a complex mixture of di- to pentapeptides, their conversion to derivatives (*O*-trimethylsilyl polyamino alcohols) that are amenable to gas chromatography and mass spectrometry, and the determination of the amino acid sequence of these peptides in a single GCMS experiment.

The use of this technique for much larger polypeptides (>200 amino acids), however, presents several potential problems. First, for large polypeptides, there is a greater chance that the small peptides identified will be repeated in the protein sequence and it will not be possible to assign these to unique blocks of codons in the DNA sequence. Because the point of correctly matching the peptide sequence identified by GCMS with the DNA sequence is to establish the phasing rigorously, it is crucial that every peptide be matched in the linear sequence to the unique site to which it corresponds.

Although one can calculate the minimum size of a peptide that is, statistically, sufficiently unlikely to be repeated in a protein of given size, such calculations fail to take account of the nonrandom occurrence of certain peptide sequences and peptide sequence repeats in real proteins. Therefore, we have examined the frequency of repeating sequences in two proteins: human serum albumin (585 residues; ref. 12) and β-galactosidase (1021 residues; ref. 13). The latter protein is the longest of known sequence and contains two regions, each about 380 amino acids long, that have a high degree of sequence homology (14). Thus, this protein would be expected to have an unusually high frequency of sequence repetition. The results of this analysis suggest that, for both proteins, about 25–35% of the tripeptide sequences are repeated (Table 1). On the other hand, 96% of the tetrapeptide sequences are unique. This is true even for β-galactosidase, because the regions of high homology generally do not exhibit exact matches over more than three consecutive residues. As would be expected, pentapeptide sequences are even less likely to be repeated. Therefore, in our subsequent experiments, we have concentrated on the generation and identification of tetra- and pentapeptides.

Second, there is the possibility that a given peptide sequence will appear to match an incorrect reading frame of the DNA. Because any given DNA sequence represents three possible reading frames, for a sequence of $M$ bases, there are two incorrect amino acid sequences, each $M/3$ amino acids long. The frequency at which any tetrapeptide will be found again by random chance in either of the incorrect reading frames is about $2 \times (M/3)/20^4$ or $M \times 4.16 \times 10^{-6}$. Therefore, for $M = 3000$ (corresponding to a chain of 1000 amino acids), the random frequency of encountering that same peptide in an incorrect reading frame a second time is ≈0.012. Even when leucine and isoleucine are considered as a single amino acid (because they are difficult to distinguish by mass spectrometry) and asparagine and glutamine are not distinguished from aspartic acid and glutamic acid, respectively (as would be the case when the peptides were esterified by using MeOH/HCl), this probability is only increased to 0.024. Thus, even for very large polypeptides, the possibility of placing a specific tetrapeptide in the wrong reading frame is very small. Furthermore, in the rare case where a peptide is placed into the wrong reading frame, this error will surely become apparent when it is found to fit elsewhere in the correct reading frame as established by the locations of other peptides.

A third potential problem is that the GCMS method requires that the peptide derivatives be transmitted through a gas chromatograph. This generally poses no problem for the de-

Table 1. Percentage of peptides that are unique

| | Human serum albumin | β-Galactosidase |
|---|---|---|
| Dipeptides | 8.6 | 5.5 |
| Tripeptides | 72.4 | 64.9 |
| Tetrapeptides | 96.6 | 95.7 |
| Pentapeptides | 99.7 | 99.0 |

Uniqueness of oligopeptides derived from human serum albumin and β-galactosidase. Leucine and isoleucine, aspartic acid and asparagine, and glutamic acid and glutamine were considered to be indistinguishable.

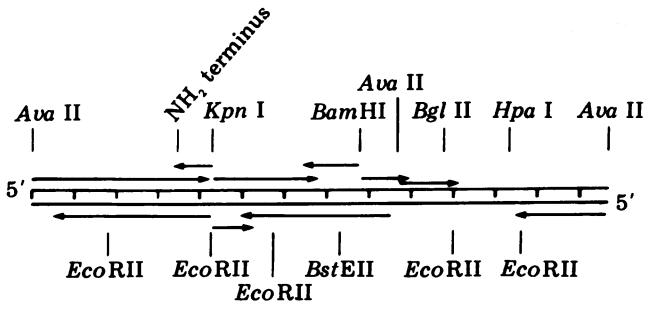


FIG. 1. Restriction map of the region of DNA corresponding to the $NH_2$-terminal half of AlaRS. Arrows delineate the lengths of the segments that were sequenced, the direction of sequencing, and the strand that was sequenced. Vertical tick marks are placed at every 100 base pairs.

rivatives of di- and tripeptides but is more difficult for certain larger ones. Therefore, we have calculated the proportion of tetra- and pentapeptides, derivable from human serum albumin and β-galactosidase, that are amenable to gas chromatography as *O*-trimethylsilyl polyamino alcohols. This calculation, which is based on the predictability of their gas chromatographic properties (5), shows that 86–91% of the tetrapeptides and 54–59% of the pentapeptides from these two proteins are amenable to gas chromatography under the conditions we currently use (5, 6).

The final problem in applying the GCMS technique to a very long polypeptide is the extreme complexity of the resulting hydrolysate. It was therefore necessary to determine whether these complex mixtures could be sufficiently resolved by gas chromatography to produce interpretable mass spectra. In model studies using bovine serum albumin ($M_r$ = 67,000), we found that, even the relatively low resolving power of packed gas chromatographic columns allowed unambigious identification of the 30–50 peptides produced in highest yield. Thus, there appears to be no serious obstacle presented by the complexity of the hydrolysate of a long polypeptide.

**Sequence of Alanine tRNA Synthetase.** *Escherichia coli* AlaRS is an $\alpha_4$ tetramer in which the subunit has a chain length of almost 900 amino acids (3). The cloning and restriction mapping analysis of the gene for it and the nucleotide sequence of (largely) the pregene region have been reported (2).

Digestion of AlaRS with trypsin produces an $NH_2$-terminal fragment (T-1) that has approximately 360 amino acids. A restriction map of the portion of the gene corresponding to the fragment is shown in Fig. 1, which indicates the regions that have been sequenced and the strand and direction of sequencing; for most of the region corresponding to fragment T-1, either one or both strands has been sequenced.

The T-1 fragment was subjected to partial digestion by both enzymatic and chemical procedures under conditions that maximize the yields of tetra- and pentapeptides. Each partial hydrolysate was converted to the *O*-trimethylsilyl polyamino alcohols and analyzed by GCMS. For example, most of the 42 peptide derivatives that have been identified in the total ionization plot for the thermolytic digest correspond to di- and tripeptides, but 7 of them correspond to tetra- and pentapeptides (Fig. 2). Similar data were obtained from GCMS analyses of other partial digests.

The sequence of the first 11 amino acids of the $NH_2$-terminus of the T-1 fragment was determined by an Edman degradation (3). In our initial DNA sequences, we were unable to find a stretch of bases, on either strand of the DNA, that corresponded to the $NH_2$ terminus. However, by using a computer, some of the tetra- and pentapeptide sequences obtained from the GCMS analysis could be fitted to unique clusters of codons in the regions of the DNA that were sequenced. In this way, the GCMS data quickly established the "sense" strand of the DNA and the proper reading frame. By knowing the sense strand and, therefore, the direction of transcription, we deduced that the $NH_2$ terminus must fall near the *Kpn* I site (see Fig. 1). DNA sequencing was then concentrated in this area, and thus the $NH_2$-terminal encoding region of *alaS* was found more easily than would have otherwise been possible.

The sequence of the sense strand of the first 500 nucleotides of *alaS* is shown in Fig. 3. In this region, both strands have been sequenced completely. The various peptides span most of the length of the sequenced DNA, and all of them fall in exactly the same reading frame. Because of the perfect phasing of each of the 9 tetrapeptides to the first 500 nucleotides, confidence

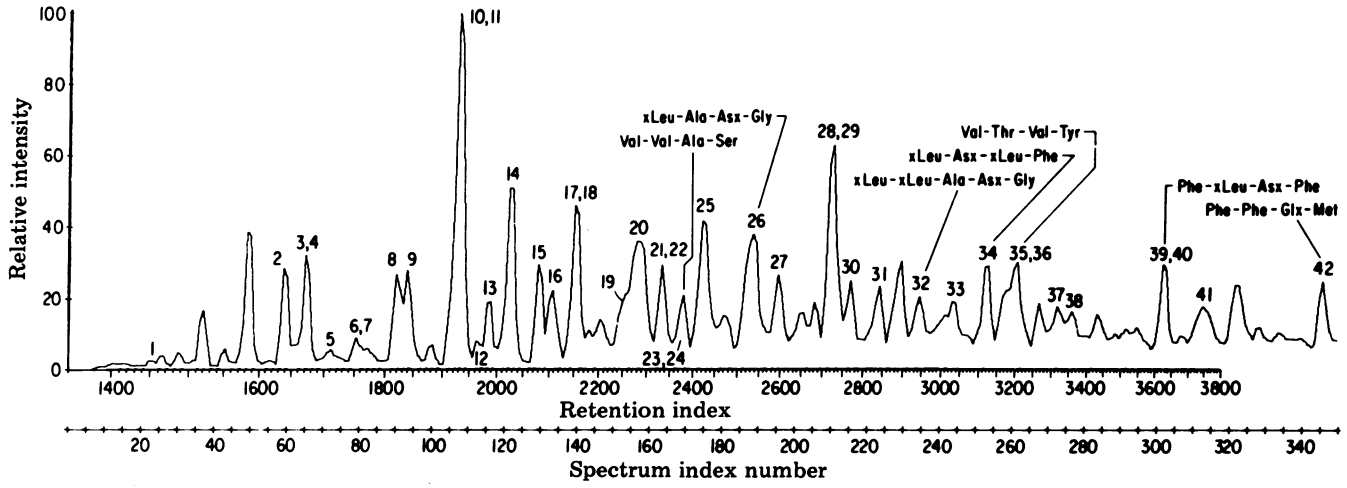


FIG. 2. Mass resolved (15) total ionization plot of thermolytic digestion product of fragment T-1. Numerals refer to peaks representing identified peptide derivatives—only the structures of tetra- and pentapeptides are shown. Aspartic acid and glutamic acid are not distinguished from asparagine and glutamine, respectively, because the amides are converted to the methyl esters in the esterification step. Leucine and isoleucine are indistinguishable in the spectra of the larger peptide derivatives. When matching these peptides to the base sequence shown in Fig. 3, all permutations of these three pairs were considered. Because they do not fit on the first 500 nucleotides in any reading frame, peptides 26, 32, and 34 must be derived from beyond amino acid position 165 in the T-1 fragment. xLeu = leucine or isoleucine, Asx = asparagine or aspartic acid, and Glx = glutamine or glulamic acid.

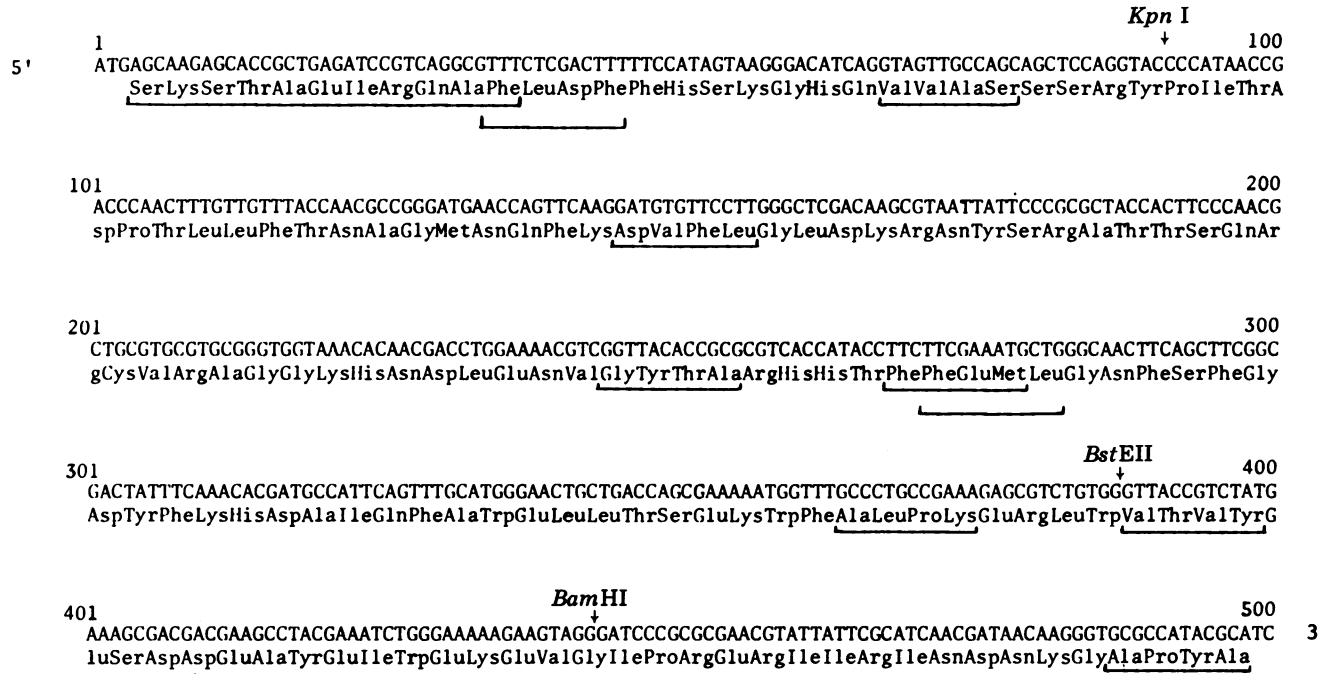


FIG. 3. Nucleotide sequence of the portion of DNA encoding the $NH_2$-terminal region of AlaRS (in this region, both strands of the DNA have been sequenced) and polypeptide sequence derived from the DNA sequence. Tetrapeptide sequences determined by GCMS analysis of protein hydrolysates of the $NH_2$-terminal T-1 fragment and the undecapeptide $NH_2$-terminal sequence determined by Edman degradation are indicated by underbars in the locations where they fit unique consecutive blocks of codons. Because the amino acid pairs Leu and Ile, Asp and Asn, and Gln and Glu are not distinguished by the GCMS analysis, the amino acid predicted by the DNA sequence is given. All of the peptides fit in the same reading frame. (Compare with Fig. 1.)

can be placed in the complete polypeptide sequence derived from the translated DNA sequence.

**Comments on Multiple-Phase Check.** There is no question that sequencing both strands of a DNA molecule is a valid, and generally accurate, way to determine polypeptide sequences. However, for very long polypeptides, corroborative data on the protein itself are essential. One cannot simply rely on encountering premature stop codons to serve as a check on the accuracy of the nucleic acid sequence that is translated into a polypeptide. For example, in AlaRS, there is a stretch of bases from position 222 to position 343 in which there are no stop codons in any reading frame (see Fig. 3). In addition, the recent discovery of the use of a stop codon (i.e., UGA) as a signal for a specific amino acid (16, 17), and other unusual features of coding (16, 18), reaffirm the need for independent corroborative data on the protein of interest.

The main point of the multiple-phase check is to ensure the accurate deduction of polypeptide sequences from data that rely heavily on DNA sequence information. Any approach that gives scattered oligopeptide sequences, preferably of tetra- or of higher order peptides (see above), is useful, including current methods of peptide separations for sequencing by Edman degradation. The particular advantage of the GCMS approach is that it can determine a large number of oligopeptide sequences from widely scattered regions of the structure in a single experiment (without isolation of pure peptides) and, furthermore, that inherent to the method, many of these sequences are of exactly the size required to ensure that they can be placed on unique blocks of codons. Thus, although the peptide sequences determined by using the GCMS approach to protein sequencing are too short to build up good overlaps, the information obtained is ideal for establishing a phase check of translated DNA sequences. And, for establishing a multiple-phase check, there is no more rapid approach. Also, although we have used a bacterial strain that overproduces the protein of interest, adequate information can be obtained from a single experiment with 2–5 mg of protein sample; thus, large amounts of protein are not required.

Of the 165 codons shown in Fig. 3, we have nine sequences of tetrapeptides scattered along the chain, which gives an average frequency of about one peptide for every 19 residues. This corresponds to a phase check on the DNA at an average interval of every 57 nucleotides. Given the accuracy of DNA sequencing, it is doubtful that a greater density of checks is required.

In summary, the combination of GCMS analysis and DNA sequencing is particularly useful in three ways: (*i*) With a minimal amount of DNA sequence information, the GCMS rapidly determines the sense strand of the DNA, which, in turn, gives the direction of transcription; this information can strongly influence the choice of regions of the DNA to sequence in subsequent experiments, because it can give a rough idea of the location of the $NH_2$-terminal coding region and can be used in conjunction with other data to define approximately the limits of the DNA that encodes the polypeptide structure. (*ii*) The GCMS analysis provides a direct determination of the polypeptide sequence in those regions of the structure from which peptides can be detected by the methodology. Therefore, these portions of the polypeptide structure are independently determined by two different approaches applied to two different molecules. (It should be noted that we have identified a large number of tripeptides that can be used to confirm additional parts of the amino acid sequence although, due to the increased chance that some of them will also occur in the COOH-terminal half of the T-1 fragment, they can be used with confidence only when the entire sequence has been completed.) (*iii*) The GCMS approach is well suited to providing phase checks on the translated DNA sequence and therefore independently establishes the proper phase of the DNA throughout its entire length.

We gratefully acknowledge the collaboration of Dr. R. J. Anderegg in the early phases of this work, the excellent technical assistance of B. Meeusen and E. Block, and the programming assistance of T. Royal. This investigation was supported by National Institutes of Health Grants GM05472 (to K.B.), GM15539, and GM23562 (to P.R.S.), S.D.P. is a National Institutes of Health Predoctoral Trainee.

1. Schimmel, P. R. & Söll, D. (1979) *Annu. Rev. Biochem.* **48,** 601–648.
2. Putney, S. D., Meléndez, D. L. & Schimmel, P. R. (1980) *J. Biol. Chem.*, in press.
3. Putney, S. D., Sauer, R. T. & Schimmel, P. R. (1980) *J. Biol. Chem.*, in press.
4. Laemmli, U. K. (1970) *Nature (London)* **227,** 680–685.
5. Nau, H. & Biemann, K. (1976) *Anal. Biochem.* **73,** 139–153.
6. Carr, S. A., Herlihy, W. C. & Biemann, K. (1980) *Biomed. Mass Spectrom.*, in press.
7. Herlihy, W. C. & Biemann, K. (1980) *Biomed. Mass Spectrom.*, in press.
8. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
9. Biemann, K. (1978) *Pure Appl. Chem.* **50,** 149–158.
10. Biemann, K. (1980) in *Biochemical Applications of Mass Spectrometry,* First Supplementary Volume, eds. Waller, G. R. & Dermer, Q. C. (Wiley, New York), pp. 469–525.
11. Khorana, H. G., Gerber, G. E., Herlihy, W. C., Gray, C. P., Anderegg, R. J., Nihei, K. & Biemann, K. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 5046–5050.
12. Meloun, B., Morávek, L. & Kostka, V. (1975) *FEBS Lett.* **58,** 134–137.
13. Fowler, A. V. & Zabin, I. (1978) *J. Biol. Chem.* **253,** 5521–5525.
14. Hood, J. M., Fowler, A. V. & Zabin, I. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 113–116.
15. Biller, J. E. & Biemann, K. (1974) *Anal. Lett.* **7,** 515–528.
16. Barrell, B. G., Bankier, A. T. & Drouin, J. (1979) *Nature (London)* **282,** 189–194.
17. Geller, A. I. & Rich, A. (1980) *Nature (London)* **283,** 41–46.
18. Barrell, B. G., Anderson, S., Bankier, A. T., DeBruijn, M. H. L., Chen, E., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 3164–3166.