Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 1956–1960, April 1984
Biochemistry

Strategy for the mass spectrometric verification and correction of the primary structures of proteins deduced from their DNA sequences

(fast atom bombardment-mass spectrometry/tryptic peptides/molecular weight determination)

BRADFORD W. GIBSON AND KLAUS BIEMANN*

Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139

Communicated by John M. Buchanan, December 5, 1983

ABSTRACT Fast atom bombardment mass spectrometry has been used to confirm and correct regions from the amino acid sequences of three large proteins, glutaminyl- and glycyl-tRNA synthetase from *Escherichia coli* and methionyl-tRNA synthetase from yeast, whose primary structures had been deduced from the base sequences of their corresponding genes. The strategy is based on a comparison of the molecular weights of the tryptic peptides predicted from all three reading frames of the gene sequences with those determined mass spectrometrically. The experimental molecular weights either match or differ and can be used to assess the correctness of the base sequences, identify errors that lead to frame shifts, premature stop codons, incorrect amino acids, etc., or identify the presence of posttranslational modifications. This method is very fast and requires little material (5–20 nmol).

The determination of the primary structure of large proteins (M_r 50,000 and above) has recently been revolutionized by the development of fast and reliable techniques for the base sequence analysis of the DNA encompassing the gene encoding the protein (1–4). They are much faster and more efficient than conventional protein sequence analysis techniques and are now the method of choice for establishing the primary structure of a protein whenever applicable. Even though the results are most likely complete and accurate if both strands of the DNA are sequenced and other precautions are followed, errors are still possible. In one case, separate laboratories independently determined the base sequence of a gene but with sufficiently different results so that large regions of the corresponding amino acid sequence differed drastically (5, 6). More recently, the primary structures of two proteins deduced from the Rous sarcoma virus gene had to be significantly modified to take into account numerous errors in the DNA sequence (7, 8). These errors were suspected by the investigators only when contradictory DNA and peptide sequence data became available.

It is obvious that it is prudent to check the correctness of the amino acid sequence derived from the base sequence of the gene not only at the NH_2 and COOH termini, which is the common practice, but throughout the entire protein. This would help to uncover any significant errors as well as address the possibility of posttranslational modifications. Furthermore, carrying out this check in parallel with the accumulation of double-stranded DNA sequence data increases the speed and reliability of the process because it eliminates the need to recheck DNA sequences that already match the protein data. For this purpose we successfully applied a mass spectrometric peptide sequence analysis technique developed earlier (9) to alanyl-tRNA synthetase (monomer, M_r 95,000) (10) and glutaminyl-tRNA synthetase (M_r 60,000) (11). The results demonstrated the ease and speed with which the type and location of base sequence analysis errors can be found.

More recently we have developed a strategy that is still faster, more thorough, and requires much less protein. It is based on a new mass spectrometric technique that allows the ionization (and thus molecular weight determination) of relatively large, nonvolatile molecules previously not amenable to mass spectrometry (12). The sample, dissolved in glycerol, is bombarded with a beam of atoms of a few kilovolts kinetic energy (fast atom bombardment or FAB). Protonated molecules (MH^+) and sometimes cationated molecules, MNa^+ and MK^+, are sputtered off and mass analyzed in the mass spectrometer (MS).

For checking the correctness of a proposed structure the protein is hydrolyzed at specific sites—i.e., with trypsin—to produce a pool of smaller peptides, which is then partially separated by HPLC into five or six fractions. Each fraction is then subjected to FAB-MS, which allows the determination of the molecular weights of most or all of the peptides present in each fraction. These values are then compared with the molecular weights of the tryptic peptides predicted from the DNA deduced amino acid sequence. The fractions can then also be subjected to Edman degradation(s) and the molecular weights of the shortened peptides determined by FAB-MS to identify the NH_2-terminal amino acid(s) of each peptide from the change in its molecular weight.

There are four situations that can result. First, the experimentally determined values agree (after correction for the attached proton or alkali ion) with the molecular weights predicted for the primary tryptic peptides as well as those arising from incomplete hydrolysis—i.e., peptides still containing an internal lysine or arginine. Second, most agree with the proposed sequence from one reading frame, while the remainder fits another reading frame, indicating a combination of deletion/insertion errors that must be located in the regions where the sequence goes out of phase and then back in again or a single such error near the NH_2 or COOH termini. Third, a majority of the expected and determined values match in one reading frame, while others do not fit any reading frame, indicating a number of possible errors or complications: (*a*) a pair of deletion and insertion errors so closely spaced that no tryptic peptide is located between them; (*b*) misidentification of one or more base(s), which changes the codon for one or more amino acid(s); (*c*) the gene product is larger than expected, due to an error near the COOH terminus that leads to a premature stop codon; (*d*) a posttranslational modification of the protein; (*e*) random (nontryptic) cleavage at sites other than lysine or arginine due to contaminating enzymes, inadvertent chemical hydrolysis, etc.; and (*f*) microheterogeneity of peptidal impurities in the protein preparation. Last, only very few or none of the experimen-

Abbreviations: FAB-MS, fast atom bombardment mass spectrometry; MH^+, protonated molecular ion.
*To whom reprint requests should be addressed.

tally determined peptide molecular weights matches the predictions, indicating that either the DNA sequence contains many errors or the sequenced region does not contain the gene that encodes for that protein.

We have applied this strategy to our earlier work on Gln-tRNA synthetase (11) and more recently to studies on glycyl- and methionyl-tRNA synthetases (13, 14). The FAB-MS data revealed a number of such errors in the DNA sequence and these were later substantiated by reinspection of the original DNA sequence data or repeat of the experiments. The published final sequences (11, 13) took these corrections into account without recounting the chronology of the preliminary data and their confirmation and correction. The primary structure of Met-tRNA synthetase had been deduced from the base sequences of its gene (14). The FAB-MS data on the protein showed it to be essentially correct except that the NH_2 terminus had been modified. It is the purpose of this paper to describe these strategies with specific examples.

MATERIALS AND METHODS

Gln-tRNA synthetase, isolated and purified from culture extracts of *Escherichia coli* K-12 (8), was a gift of Dieter Soll (Yale University). Gly-tRNA synthetase, also from *E. coli*, was obtained from T. Webster and P. Schimmel (Department of Biology, Massachusetts Institute of Technology). Carboxymethyl Met-tRNA synthetase from yeast was provided by Y. Boulanger and F. Fasiolo (Institut de Biologie Moleculaire et Cellulaire, Strasbourg, France). Trypsin (treated with diphenylcarbamoyl chloride) was from Sigma. Glycerol was vacuum distilled prior to use. Trifluoroacetic acid, phenylisothiocyanate, *n*-heptane, and pyridine were from Pierce for Edman sequence analysis. All other chemicals were of reagent grade.

Approximately 20 nmol each of carboxymethyl Gln- and Met-tRNA synthetases and heat-denatured Gly-tRNA synthetase were separately dissolved in 0.5–1.0 ml of 50 mM ammonium acetate buffer (pH 8.5) containing 0.5 mM $CaCl_2$. The solutions were incubated with trypsin (enzyme/substrate ratio of 1:75) at 37°C for 4 hr. The progress of the digestion was monitored by HPLC (15). The lyophilized hydrolysates were partially fractionated by HPLC on a reverse-phase C_{18} column (Waters Associates) using 0.1% aqueous trifluoroacetic acid and 0.07% trifluoroacetic acid in acetonitrile. Five or six fractions (6–8 ml each) were collected during a linear solvent program (0–60% acetonitrile). The eluant was monitored at 215 nm to insure proper collection and to minimize peak splitting between fractions. The fractions were then dried and dissolved in ≈10 μl of 50% acetic acid and 15 μl of glycerol.

A Varian MAT 731 mass spectrometer, fitted with a modified Ion Tech FAB gun operated with xenon, was used (16). Approximately 0.5–1.0 μl of each sample was placed on a stainless steel probe and inserted into the ion source. Mass spectra were recorded at a scan speed of 512 sec per decade over a mass range of 1–1960 at 8 kV accelerating voltage.

For manual Edman degradation, the material remaining from the first mass spectrometric experiments was separated from the glycerol by loading it in 0.5% acetic acid onto a Sep-Pak C_{18} cartridge (Waters Associates). The peptides were eluted with acetonitrile/water (1:1) containing 0.5% acetic acid, dried, and subjected to the manual Edman degradation for short apolar peptides (17). Mass spectra of these shortened peptides were then obtained as described above.

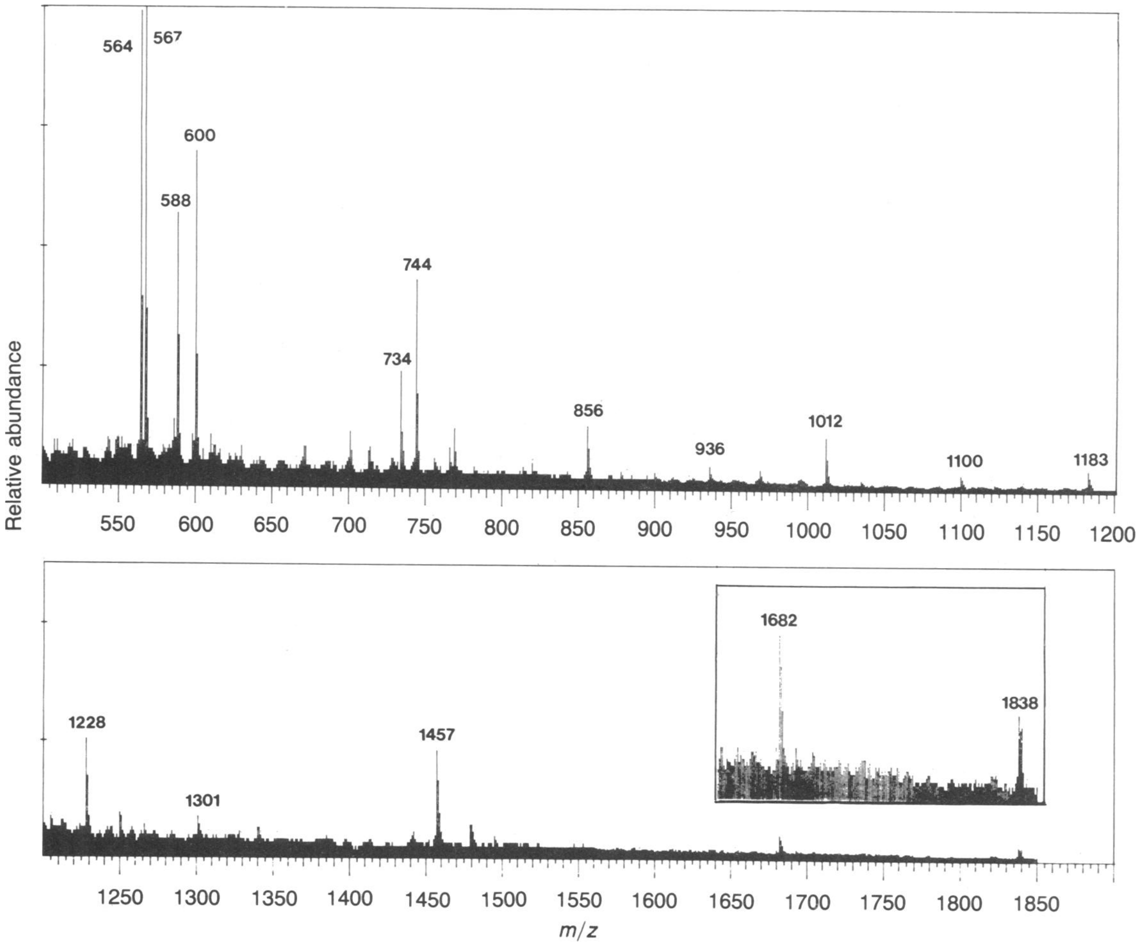

FIG. 1. FAB-MS of HPLC fraction 2 from the tryptic hydrolysate of Gly-tRNA synthetase. The MH^+ ions for each peptide are labeled.

Table 1. Peptides identified by their molecular weights in HPLC fraction 2 of tryptic digest of Gly-tRNA synthetase including two consecutive manual Edman cycles

			First Edman cycle			Second Edman cycle		
MH^+	Subunit reading frame	Position	MH^+	Position	Amino acid removed	MH^+	Position	Amino acid removed
444	β1	52–55	466	53–55*	Leu			
564	α1	272–275	401	273–275	Tyr	288	274–275	Ile
567	β1	140–144	589	141–144*	Leu	492	142–144*	Pro
588	β1	387–392	652†	388–392*	Ala	595	389–392*	Gly
600	β1	51–55	579†	52–55*	Arg			
734	No match		756		Leu (Lys*)	655		Thr
744	β1	584–590	780†	585–590	Val	693	586–590	Ser
	β1	606–611 (Glu-Pro- . . .)						
856	β3	340–346	743	341–346	Leu			
936	β1	111–117 (Gly-Glu- . . .)						
	No match		972		Val (Lys*)	841		Met
1012	β3	339–346	856	340–346	Arg	743	341–346	Leu
1100	β1	282–291	1029	283–291	Ala	930	284–291	Val
1183	β1	56–66	1084	57–66	Val	1013	58–66	Ala
1228	β1	309–319	1248	310–319*	Asp	1151	311–319*	Pro
1301	β1	551–561	1145	552–561	Arg	1048	553–561	Pro
1457	β1	599–611	1628†	600–611‡	Val			
1682	β1	56–70	1718†	57–70*	Val			
1838	β1	56–71						

*Contains a lysine that reacted with phenylisothiocyanate, adding 135 daltons.

†Accompanied by a peak 135 mass units lower, corresponding to the peptide in which ε-NH_2 of lysine has not reacted with phenylisothiocyanate.

‡Contains two lysine that reacted with phenylisothiocyanate, adding 270 daltons.

RESULTS AND DISCUSSION

In the course of the determination of the structure of Gln-tRNA synthetase we became reasonably confident in the general correctness of the amino acid sequence as deduced from the base sequence of its corresponding gene and cross-checked by the GC/MS method (11). This gene–protein pair thus provided a test case for the strategy outlined above.

A tryptic digest of this protein was separated into six crude fractions by HPLC to remove the enzyme, reagents, salts, and other contaminants and also to reduce the complexity of the original digest, which, in turn, reduces the complexity of the resulting FAB mass spectra. In five of the six HPLC fractions a total of 34 peaks could be identified as MH^+s of peptides.

When these were matched with the molecular weights of the tryptic peptides predicted from the proposed amino acid sequence of Gln-tRNA synthetase, 30 of them agreed, confirming 42% of the sequence, which indicated the high degree of correctness of the deduced structure. There was no indication of a frame shift problem, but some of the nonmatches would be due to an incorrect amino acid caused by a misidentified base. This would lead to a peptide that differs in molecular weight from a predicted one by the mass difference of two amino acids. Of course, the amino acid that is to be replaced has to be present in the DNA-derived sequence and the corresponding codons must be interconvertible, preferably by a single base change.

Misidentified Bases. There were two predicted tryptic peptides of M_rs 736 and 1379, respectively, which is 10 daltons less than two unmatched experimentally found peptides (M_rs 746 and 1389). The former cover the amino acid positions 7–12 and 1–12—i.e., they are overlapping peptides from the NH_2 terminus of Gln-tRNA synthetase. Only the pair serine/proline differs by 10 daltons. The preliminary sequence indeed contained serine in position 7, which is common to both, and the codon used for serine was TCG, which could be converted to proline codon CCG if base 19 were C rather than T.

Reinspection of the DNA sequence analysis data revealed that not only base 19 is C but also base 18. Because both CGT and CGC code for arginine, the corrected sequence now contains an arginine-proline bond, which trypsin splits very slowly (18). This is probably the reason why both tryptic peptides (positions 1–12 and 7–12) were observed. The following are the preliminary (19) and final (11) DNA and amino acid sequences of this region:

AGT-GAG-GCA-GAA-GCC-CGT-TCG-ACT-AAC-TTT -ATC-CGT
Ser - Glu - Ala - Glu - Ala - Arg - Ser - Thr - Asn - Phe - Ile - Arg
AGT-GAG-GCA-GAA-GCC-CGC-CCG-ACT-AAC-TTT -ATC-CGT
Ser - Glu - Ala - Glu - Ala - Arg - Pro - Thr - Asn - Phe - Ile - Arg

Phase Shifts. We then applied this new technique to our ongoing work on the tetrameric protein ($\alpha_2\beta_2$) Gly-tRNA synthetase (13). The α subunit has a M_r of ≈35,000 and the β subunit has a M_r of ≈65,000. The denatured protein was subjected to tryptic digestion without separation into the two subunits, which would have been a time- and material-consuming process and was unnecessary for our approach. Fig. 1 shows the FAB-MS of HPLC fraction 2 of the resulting peptide mixtures. In the spectra of the six fractions, a total of 64 signals were observed that could be assigned to MH^+ ions of peptides. Those from HPLC fraction 2 are listed in Table 1. All but 4 matched regions of the α or β subunit in reading frame 1, while 2 matched only in reading frame 3 of the β subunit and 1 (MH^+ 936) appeared to match position 111–117† of the β subunit. However, the MH^+ data after one and two Edman cycles (Table 1) indicated that this peptide begins with Val-Met . . . rather than Gly-Glu . . ., the predict-

†The numbering of the amino acid sequence of Gly-tRNA synthetase corresponds to that of the processed protein and the base sequence is numbered accordingly.

```
RF1---Glu Lys||Val Val Arg Pro Arg|Leu Ala Met Pro Ser Ser Ser Ser Thr Pro Thr Val Lys Asn------------------------------
      GAG AAA GTC GTT CGT CCG CGT CTG GCG ATG CCG AGT TCT TCT TCA ACA CCG ACC GTA AAA AAC GTC TTG AAG ATA ACC TGC CGC GCC TGC
RF3----------------------------Ser Gly Asp Ala Glu Phe Phe Phe Asn Thr Asp Arg Lys Lys |Arg |Leu Gly Asp Asn Leu Pro Arg||Leu

Base Inserted at Position  1   TCT GGC GAT GCC GAG TTC TTC TTC AAC ACC GAC CGT AAA AAA
                               |Ser Gly Asp Ala Glu Phe Phe Phe Asn Thr Asp Arg| Lys |Lys   MW= 1404, 1532

                           2   CCT GGC GAT GCC GAG TTC TTC TTC AAC ACC GAC CGT AAA AAA
                               |Pro Gly Asp Ala Glu Phe Phe Phe Asn Thr Asp Arg| Lys |Lys   MW= 1414, 1542
                           .    .   .   .   .   .   .   .   .   .   .   .   .   .   .
                           .    .   .   .   .   .   .   .   .   .   .   .   .   .   .

                           6   CTG GCC GAT GCC GAG TTC TTC TTC AAC ACC GAC CGT AAA AAA
                       →   7   CTG GCG GAT GCC GAG TTC TTC TTC AAC ACC GAC CGT AAA AAA
                               |Leu Ala Asp Ala Glu Phe Phe Phe Asn Thr Asp Arg| Lys |Lys   MW= 1444, 1572
                           .    .   .   .   .   .   .   .   .   .   .   .   .   .   .
                           .    .   .   .   .   .   .   .   .   .   .   .   .   .   .

                          15   CTG GCG ATG CCG AGG TTC TTC TTC AAC ACC GAC CGT AAA AAA
                               |Leu Ala Met Pro Arg||Phe Phe Phe Asn Thr Asp Arg| Lys |Lys   MW= 586, 945, 1073

                          16   CTG GCG ATG CCG AGT TTC TTC TTC AAC ACC GAC CGT AAA AAA
                          17   CTG GCG ATG CCG AGT TTC TTC TTC AAC ACC GAC CGT AAA AAA
                               |Leu Ala Met Pro Ser Phe Phe Phe Asn Thr Asp Arg| Lys |Lys   MW= 1444, 1572
                           .    .   .   .   .   .   .   .   .   .   .   .   .   .   .
                           .    .   .   .   .   .   .   .   .   .   .   .   .   .

                          42   CTG GCG ATG CCG AGT TCT TCT TCA ACA CCG ACC GTA AAA AAA
                               |Leu Ala Met Pro Ser Ser Ser Ser Thr Pro Thr Val Lys| Lys|   MW= 1304, 1432
```

FIG. 2. The base sequences and corresponding amino acid sequences of a few of the 42 possibilities generated by the insertion of one base (underlined) at all possible positions. The predicted tryptic peptides are underlined and their molecular weights (MW) are listed to the right (those found underlined). Amino acids identified by Edman degradation are shown as half arrows (⇀). RF, reading frame.

ed amino acids in positions 111 and 112. The Edman data also indicated that the peptide of M_r 743 (MH^+ 744) represents the sequence 584–590 rather than 606–611 in the β subunit, even though both have the same molecular weight.

From the data shown in Table 1 and those obtained from the other HPLC fractions, it is clear that there must be a pair of deletion/insertion errors between amino acid positions 324 and 387:

```
Reading frame 1                         Reading frame 1
   320---324                                387---392
              41 Bases           79 Bases
             (1 missing)         (1 extra)
                  339--346 347----360
                   Reading frame 3
```

To find the position and nature of the missing base, each of four bases would have to be inserted consecutively, one at a time, generating 168 possibilities for the 42 base sequences. The predicted tryptic peptides of the corresponding amino acid sequences would then have to be checked against those actually detected by FAB-MS. To simplify the task, it was assumed that the most likely error is the failure to recognize the correct number of consecutive identical bases (1 instead of 2 or 2 instead of 3, etc.). Thus, doubling up each base one at a time generates 42 possibilities. This is illustrated in Fig. 2 where only 7 of these are shown to save space. Of all predicted tryptic peptides, only those of M_r 1444 match a MH^+ ion actually detected (m/z = 1445 in HPLC fraction 4). These two were distinguished by three successive Edman steps that shifted the MH^+ peak to 1331, 1260, and 1145, corresponding to the consecutive removal of leucine (or isoleucine), alanine, and asparagine (but not methionine), eliminating sequences 16 and 17 but not sequences 6 and 7. Thus, either C must have been missed in position 6 or G in position 7 (counting from the last unambiguous base). Reinspection

```
631                                                                                                       657
CGT TAC CAG GAT GCG CTG GTC GAA CTG GCT GAG CTG CGT GAA CCG GTT GAT GCT TTC TTC GAT GAA AGT GAT GGT CAT GGT TGA TGA
Arg||Tyr Gln Asp Ala Leu Val Glu Leu Ala Glu Leu Arg|Glu Pro Val Asp Ala Phe Phe Asp Glu Ser Asp Gly His Gly STOPSTOP

631                                                                                                             660
CGT TAC CAG GAT GCG CTG GTC GAA CTG GCT GAG CTG CGT GAA CCG GTT GAT GCT TTC TTC GAT GAA AGT GAT GGT CAT GGT TGA TGA CAA
Arg||Tyr Gln Asp Ala Leu Val Glu Leu Ala Glu Leu Arg| ← RF 1
                                            RF 2 → Asn Arg Leu Met Leu Ser Ser Met Lys|Val Met Val Met Val Asp Asp Lys|
661
AGA ATT GCG TAT CAA CCG TCA GAC CAT GCT GGA CAA ACT GCG CG. ... ... ...

 Glu Leu Arg|Ile Asn Arg Gln Thr Met Leu Glu Lys Leu Arg ... ... ... ← RF 2

631                                                                                                             660
CGT TAC CAG GAT GCG CTG GTC GAA CTG GCT GAG CTG CGT GAA CCG GTT GAT GCT TTC TTC GAT AAA GTG ATG GTC ATG GTT GAT GAC AAA
Arg||Tyr Gln Asp Ala Leu Val Glu Leu Ala Glu Leu Arg||Glu Pro Val Asp Ala Phe Phe Asp Lys||Val Met Val Met Val Asp Asp Lys |
661                                                               680
GAA TTG CGT ATC AAC CGT CTG ACC ATG CTG GAG AAA CTG CGC GAA CTG TTC CTG CGC GTT GCG GAT ATT TCG CTG TTG CAA TAA TAA
Glu Leu Arg| Ile Asn Arg|Leu Thr Met Leu Glu Lys|Leu Arg|Glu Leu Phe Leu Arg||Val Ala Asp Ile Ser Leu Leu Gln|STOPSTOP
```

FIG. 3. (*Top*) The COOH terminus of the preliminary base sequence of the Gly-tRNA synthetase gene and the deduced amino acid sequence. (*Middle*) Extended base sequence. (*Bottom*) The corrected COOH terminus. Peptides identified by FAB-MS (⊔) and amino acids identified by Edman degradation (⇀) are indicated.

```
                START OF TRANSCRIPTION:    CTGATATGCATCTTTGTGGAGGAATAAAACTTGAAAAAAAAGCGGAAATTTACAACAAGC
                                                                                                            30
ATG TCT TTC CTC ATT TCC TTT GAT AAA TCA AAG AAA CAT CCT GCC CAT TTG CAG TTA GCG AAC AAT TTG AAG ATT GCC CTA GCA CTT GAA
Met|Ser Phe Leu Ile Ser Phe Asp Lys|Ser Lys|Lys|His Pro Ala His Leu Gln Leu Ala Asn Asn Leu Lys||Ile Ala Leu Ala Leu Glu
                                                                                                            60
TAT GCA AGC AAA AAT TTA AAG CCC GAA GTT GAC AAT GAT AAT GCT GCC ATG GAA TTG CGC AAT ACA AAG GAA CCT TTC CTT TTG TTT GAT
Tyr Ala Ser Lys||Asn Leu Lys Pro Glu Val Asp Asn Asp Asn Ala Ala|Met|Glu Leu Arg Asn Thr Lys||Glu Pro Phe Leu Leu Phe Asp
                                                                                                            89
GCT AAC GCT ATT CTA AGA TAT GTC ATG GAT TTT GAA GGT CAA ACT TCC GAC AAG TAT CAA TTT GCA TTG GCA TCT TTA CAA AAC CTG....
Ala Asn Ala Ile Leu Arg||Tyr Val|Met|Asp Phe Glu Gly Gln Thr Ser Asp Lys||Tyr Gln Phe Ala Leu Ala Ser Leu Gln Asn Leu....
```

FIG. 4. NH_2-terminal region of Met-tRNA synthetase showing the three possible initiation codons (ATG/Met) in boxes. Peptides identified by FAB-MS are underlined (⌴).

of the sequencing gel revealed that what had appeared to be a single G at position 966 was actually G-G. The additional base that caused a return to reading frame 1 between amino acids 360 and 387 was found by reinspection of the DNA sequence analysis experiment spanning that region.

Premature Stop Codons. At this point there were still a number of tryptic peptides that could not be matched with those predicted from the DNA sequence that had indicated two stop codons after amino acid 657 (see Fig. 3 *Top*). Furthermore, the predicted COOH-terminal peptide (amino acids 644–657) of M_r 1520 could not be found in any of the HPLC fractions. However, when predicting the tryptic peptides derived from an extended DNA sequence beyond the suspected COOH terminus, two more peptides matched in reading frame 2 (Fig. 3 *Middle*), indicating an insertion error between amino acids 643 and 653, which was quickly found to be the G at position 1954. When this was corrected and the base sequence further extended, a new pair of stop codons was encountered after amino acid 687 and two more peptides could be matched, including the one from the COOH terminus (Fig. 3 *Bottom*). The complete sequence of Gly-tRNA synthetase, which takes all of these findings into account, is published elsewhere (13).

Posttranslational Modification. The gene of cytoplasmic Met-tRNA synthetase from yeast had been subjected to sequence analysis by Walter *et al.* (14), but it was not possible to firmly establish at which of the first three methionine codons translation began as the protein was not amenable to Edman degradation, indicating a blocked NH_2 terminus.

A single set of FAB-MS experiments on 20 nmol of Met-tRNA synthetase revealed the presence of four peptides that could be matched with regions of the amino acid sequence between the first and second methionine, establishing that translation must have begun at the first (Fig. 4). The data also showed the presence of a tryptic peptide of M_r 997 based on a MH^+ ion at m/z = 998. This is 42 daltons (CH_3CO minus H) more than the predicted NH_2-terminal peptide without methionine, indicating that it had been removed and that the new NH_2 terminus, Ser, had been acetylated. Accurate mass measurements that gave a value of 998.5214 (duplicate, 998.5205) in very good agreement with that calculated (998.5199) for the protonated NH_2-terminal tryptic peptide CH_3CO-Ser-Phe-Leu-Ile-Ser-Phe-Asp-Lys-OH further confirmed this assignment.

Conclusions. The FAB-MS data of the mixture of tryptic peptides obtained by specific cleavage of even quite large proteins (in the case of Gly-tRNA synthetase, a mixture of two subunits) greatly increase the speed by which a reliable amino acid sequence can be derived from the base sequence of the gene which encodes it. It appears to be most useful at the point when the preliminary DNA sequence is in hand but before carrying out the redundant base sequence analysis that would be necessary in the absence of confirming amino acid sequence data. This is particularly important because the repeat experiments may be subject to the same errors that occurred the first time, which is obviously not possible for data derived from the protein.

It should be noted that not all of the peptides that would be expected have been identified in these experiments. There are two reasons for this, one of which has already been mentioned: the limited mass range of our present instrumentation, which precludes the reliable detection of peptides of M_r >2000 (commercial mass spectrometers have recently been introduced with an extended mass range to 10,000 daltons). The second reason is that not all peptides ionize equally well under FAB conditions, especially di- and tripeptides, and in mixtures competitive suppression of ionization is sometimes observed. We hope to overcome both of these limitations in the near future to achieve full coverage of large proteins.

We thank Mr. Steven Martin for modifying our mass spectrometer for FAB-MS, Drs. Hubert Scoble and Walter Herlihy for developing computer programs based on the strategies discussed in this paper, and Drs. Nancy Royal and Teresa Webster for discussions and advice on DNA sequence analysis problems. The DNA sequence data and protein samples were provided by Profs. D. Soll (Gln-tRNA synthetase), P. Schimmel (Gly-tRNA synthetase), and Y. Boulanger (Met-tRNA synthetase). This work was supported by grants from the National Institutes of Health (GM05472 and RR00317).

1. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
2. Sanger, F. & Coulson, A. R. (1975) *J. Mol. Biol.* **94,** 441–448.
3. Sanger, F., Coulson, A. R., Barrel, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143,** 161–178.
4. Messing, J., Crea, R. & Seenburg, P. H. (1981) *Nucleic Acids Res.* **9,** 309–321.
5. Heffron, F. & McCarthy, B. J. (1979) *Cell* **18,** 1153–1163.
6. Chou, J., Casadaban, M. J., Lemaux, P. G. & Cohen, S. N. (1979) *Genetics,* **76,** 4020–4024.
7. Czernilofsky, A. D., Levinson, A. D., Varmus, H. E., Bishop, J. M., Tischer, E. & Goodman, H. M. (1980) *Nature (London)* **287,** 198–203.
8. Czernilofsky, A. D., Levinson, A. D., Varmus, H. E., Bishop, J. M., Tischer, E. & Goodman, H. M. (1983) *Nature (London)* **301,** 736–738.
9. Carr, S. A., Herlihy, W. C. & Biemann, K. (1981) *Biomed. Mass Spectrom.* **8,** 51–56.
10. Putney, S. D., Royal, N. R., de Vegvar, H. N., Herlihy, W. C., Biemann, K. & Schimmel, P. (1981) *Science* **213,** 1497–1501.
11. Hoben, P., Royal, N. R., Cheung, A., Yamao, F., Biemann, K. & Soll, D. (1982) *J. Biol. Chem.* **257,** 11644–11650.
12. Barber, M., Bordoli, R. S., Sedgwick, R. D. & Tyler, A. N. (1981) *J. Chem. Soc. Chem. Commun.* **7,** 325–327.
13. Webster, T. A., Gibson, B. W., Keng, T., Biemann, K. & Schimmel, P. (1983) *J. Biol. Chem.* **258,** 10637–10641.
14. Walter, P., Gangloff, J., Bonnet, J., Boulanger, Y., Ebel, J.-P. & Fasiolo, F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 2437–2441.
15. Fullmer, C. S. & Wasserman, R. H. (1979) *J. Biol. Chem.* **254,** 7208–7212.
16. Martin, S. A., Costello, C. E. & Biemann, K. (1982) *Anal. Chem.* **54,** 2362–2368.
17. Tarr, G. E. (1977) *Methods Enzymol.* **47,** 335–357.
18. Carnegie, P. (1969) *Nature (London)* **223,** 958–959.
19. Biemann, K. (1982) *Int. J. Mass Spectrom. Ion Phys.* **45,** 183–194.