Racemization of the 2a Chiral Center in Methyl Bacteriopheophorbide *e* (2). Methyl bacteriopheophorbide *e* (**2**) (15 mg) was dissolved in trifluoroacetic acid–water (9 mL/2 mL), and sulfuric acid (1 mL) was added. This mixture was stirred at 50 °C for 1 h under nitrogen. The reaction mixture was then diluted with methylene chloride and slowly poured into aqueous saturated sodium bicarbonate. The organic layer was further washed with bicarbonate and water and was dried over anhydrous sodium sulfate. The crude dried product was treated with excess ethereal diazomethane and then purified on alumina (Brockmann grade V; elution with methylene chloride). Vis (relative absorbance): 444 nm (1.00), 538 (0.199), 574 (0.182), 604 (0.181), 660 (0.268). NMR (360 MHz, $CDCl_3$): the meso proton region of this spectrum is shown in Figure 3B. The remaining assignments were entirely consistent with Bpheo *e* given above.

Ethanedithiol Ketal (6) of Methyl Bacteriopheophorbide *e*. Methyl bacteriopheophorbide *e* (**2**) (50 mg) was dissolved in chloroform (5 mL) and cooled to −15 °C (ethylene glycol/dry ice). Ethanedithiol (8 μL, 1.1 equiv) was added via a syringe followed by boron trifluoride etherate (10 μL, 1.1 equiv), which was also added dropwise by syringe. The reaction mixture was then allowed to warm to room temperature and stir overnight under nitrogen. With no apparent change in the visible absorption spectrum, the mixture was again cooled to −15 °C, 1 additional equiv of boron trifluoride etherate (10 μL) was added, and the reaction mixture was again stirred overnight under nitrogen. At this time spectrophotometry showed a dramatic hypsochromic shift of the Soret band from 444 to 426 nm. The reaction mixture was diluted with methylene chloride, washed with water, dried, and evaporated to dryness. Vis (relative absorbance): 426 nm (1.00), 528 (0.078), 562 (0.093), 608 (0.0739), 664 (0.286). IR ν(C=O) (relative absorbance), 1733 (0.56, ester), 1694 cm^{-1} (1.00, 9-keto). MS, *m/e* (%), $(M + 29)^+$ 713 (36), 727 (10), 741 (5); $(M + H)^+$ 685 (42), 699 (40), 713 (36); $[(M + H) - 18]^+$ 667 (55), 681 (40), 695 (25). NMR (360 MHz, $CDCl_3$): 10.92, 10.88, 10.85, 10.83, 10.81 (s, α-meso H); 9.54, 9.52, 9.49 (s, β-meso H); 6.95, 6.91, 6.88 (s, 3a-CH); 6.50 (m, 2a-CH); 5.22 (m, 10-CH_2); 4.58 (q, 8-H); 4.25–4.05 (m, 7-H, 4a- and 5a-CH_2); 3.85 (s, δ-meso Me); 3.78 (m, 3-SCH_2CH_2S); 3.60 (s, 7d-OMe); 3.50, 3.51 (s, 1-Me); 2.83 (broad s, 2a-OH); 2.50, 2.20 (m, 7-CH_2CH_2, 4b-CH_2, 4c-CH, 2b-Me); 1.95 (t, 5b-Me); 1.75 (t, 4b-Me); 1.50, 1.25, 0.85 (m, 8-Me, 4c-Me, 4d-Me); −1.60 ppm (broad m, NH).

Methyl Bacteriopheophorbide *c*. The dithioacetal derivative **6** (27 mg) was dissolved in methanol–tetrahydrofuran (30 mL/5 mL), to which was added Raney nickel (540 mg, 20 mass equiv) in a pH 10 slurry. This flask was sealed and heated while stirring for 1 h in the dark. The Raney nickel was filtered off on Celite and the solution was evaporated to dryness. The dried product was dissolved in methylene chloride (50 mL) and washed with water (2 × 50 mL), dried, and again evaporated to dryness. The product was purified by preparative TLC on silica gel (elution with 3% methanol–methylene chloride) and obtained as a solid from methylene chloride–*n*-hexane. Vis (relative absorbance): 414 nm (1.00), 520 (0.110), 552 (0.152), 612 (0.100), 670 (0.457). IR ν(C=O) (relative absorbance), 1735 (0.56, ester), 1689 cm^{-1} (1.00, 9-keto). MS, *m/e* (%), $(M + 29)^+$ 623 (60), 637 (13), 651 (10); $(M + H)^+$ 595 (85), 609 (100), 623 (60); $[(M + H) - 18]^+$ 577 (15), 591 (18), 605 (10). NMR (360 MHz, $CDCl_3$): 9.96, 9.95, 9.94 (s, α-meso H); 9.54, 9.52, 9.50 (s, β-meso H); 6.56 (m, 2a-CH); 5.26 (m, 10-CH_2); 4.59 (q, 8-H); 4.21 (d, 7-H); 4.12, 3.72 (m, 4a- and 5a-CH_2); 3.90 (s, δ-meso Me); 3.58 (s, 7d-OMe); 3.54 (s, 1-Me); 3.30 (s, 3-Me); 2.61 (broad s, 2a-OH); 2.54, 2.18 (m, 7-CH_2CH_2, 4b-CH_2, 4c-CH, 2b-Me); 1.96 (t, 5b-Me); 1.71 (t, 4b-Me); 1.50, 1.22 (m, 8-Me, 4c-Me, 4d-Me); −1.76, −1.78 ppm (broad s, N H).

Acknowledgment. This research was supported by a grant from the National Science Foundation (CHE-81-20891). We are grateful to Professor Norbert Pfennig (Konstanz) for advice and assistance, as well as for provision of Bchl *e* cultures. We also thank Dr. Daniel J. Repeta and Mr. Carl G. Johnson (W.H.O.I.) for assistance in obtaining mass and IR spectra.

Rapid Protein Sequencing by the Enzyme–Thermospray LC/MS Method

Krystyna Stachowiak, Cheryl Wilder, Marvin L. Vestal, and Douglas F. Dyckes*

Contribution from the Department of Chemistry, University of Houston, Houston, Texas 77004.

Received June 9, 1987

Abstract: A rapid on-line system for the analysis of protein sequences has been developed. This system uses a combination of immobilized enzymes, liquid chromatography columns, and a thermospray LC/MS. A denatured protein injected into this system is subjected to endopeptidase proteolysis, and the resultant fragments are separated chromatographically and analyzed by mass spectrometry, all in a continuous process. Sequence information on the endopeptidase fragments is generated by inserting a column of immobilized exopeptidase on-line, following the LC column. This causes the fragment in each chromatographic peak to be further digested as it emerges from the LC column and results in a set of sequence peptides. These peptides, which result from the loss of one, two, or more terminal amino acid residues, can all be logically related to the original fragment, and the partial sequence deduced. Through multiple combinations of endopeptidases, exopeptidases, and LC columns, it is possible to determine large portions of the protein sequence. This method is demonstrated for basic pancreatic trypsin inhibitor as the substrate and trypsin, chymotrypsin, aminopeptidase M, and carboxypeptidases B and Y as the immobilized peptidases. The positions of all of the tryptic and chymotryptic fragments of the trypsin inhibitor can be deduced, and fully one-half of the amino acid residues can be assigned to their correct positions on the basis of a series of ten experiments, none of which require over an hour of instrument time.

The determination of the sequence of a protein is the first step in a comprehensive understanding of how it works. The amount of information required and the frequent scarcity of material have made protein sequence determination a continuing challenge.

Edman's automation[1] of his classical chemical sequencing approach[2] was the first major step toward the development of rapid protein sequencing. The Edman method has now been extended to the picomolar level[3] and remains the principal technique for quick, sensitive, and reliable determination of sequence information from proteins. An entirely different approach, of sequencing proteins indirectly, through sequencing their genes,[4,5] has made it possible to generate massive amounts of primary sequence data,

[1] Edman, P.; Begg, G. *Eur. J. Biochem.* **1967**, *1*, 80.
[2] Edman, P. *Acta Chem. Scand.* **1950**, *4*, 283.
[3] Hedwick, R. M.; Hunkapiller, M. W.; Hood, L. E.; Dryer, W. J. *J. Biol. Chem.* **1981**, *256*, 7790.
[4] Gilbert, W. *Science (Washington, D.C.)* **1981**, *214*, 1305.
[5] Sanger, F. *Science (Washington, D.C.)* **1981**, *214*, 1205.

0002-7863/88/1510-1758$01.50/0

even for proteins that have not yet been isolated. Despite the power of these approaches, neither can solve every problem in the primary structure of proteins, and research on alternative approaches to sequencing continues to flourish. Among the alternatives, mass spectrometric methods stand out.

The analysis of protein structure by mass spectrometry (MS) was advanced significantly by the discovery of methods such as ^{252}Cf plasma desorption,[6,7] fast atom bombardment,[8] and thermospray,[9] all of which permit the ionization of underivatized peptides and proteins. The combination of such ionization techniques with tandem MS for the analysis of protein structure[10,11] has been shown to be a powerful technique, even applicable for the analysis of structural modifications of individual peptides in a peptide mixture.[12] An alternative approach, which uses enzymatic cleavage of proteins and liquid chromatographic (LC) separation of the components prior to thermospray MS analysis, is described below.

Thermospray LC/MS permits the direct analysis of substances in an LC eluate.[9,13,14] The eluate, normally an aqueous solution containing a volatile buffer, is introduced to the mass spectrometer through a heated capillary, from which it emerges as a supersonic jet containing a mixture of vaporized solvent and fine droplets. This high-velocity stream passes through a heated ion source into a vacuum line. Ions are evaporated from the droplets' surfaces as they traverse the ion source, and diffuse through an aperture into the mass analyzer. Using this technique it is possible to detect ions of nonvolatile samples. For example, it is possible to detect dissolved, underivatized peptides of masses up to at least 4000 amu, without significant fragmentation.[15] This instrumental advance has permitted us to develop methods for the analysis of peptides by LC/MS.[15-18]

A continuous-flow system for the rapid sequence analysis of peptides has been reported in detail.[18] This system consisted of a series of analytical columns connected on-line and terminating with the thermospray LC/MS. First, a column of immobilized trypsin was used to cleave the peptide being analyzed into primary fragments. Next, an Ultrasil-NH_2 LC column separated these primary fragments. A second enzyme column, of immobilized carboxypeptidase Y (CPY), then digested each fragment as it emerged from the Ultrasil column. Finally, the thermospray mass spectrometer served as the LC/MS detector. A block diagram of the system is shown in Figure 1.

The analysis of a peptide by this enzyme–thermospray LC/MS system was demonstrated using α-MSH. The peptide was split into two primary fragments, one of eight and one of five amino acid residues, upon passage through the trypsin column. The two fragments were fully resolved by Ultrasil, and each was digested during its subsequent passage through the CPY column to produce a set of sequence peptides. The sequence peptides from each primary fragment emerged simultaneously (as a single peak in the total ion chromatogram) but were detected as their individual mass ions by the thermospray mass spectrometer. Thus the mass spectrum for each chromatographic peak consisted of the parent primary fragment and one or more sequence peptides, corresponding to the removal of one, two, or more amino acid residues from the C terminus of the primary fragment. Partial sequences of each primary fragment could be deduced directly from the mass differences of the successive sequence fragment ions. Seven of the thirteen amino acid residues of α-MSH were assigned in a single analysis, which could be completed on 5 nmol of peptide in less than 5 min.

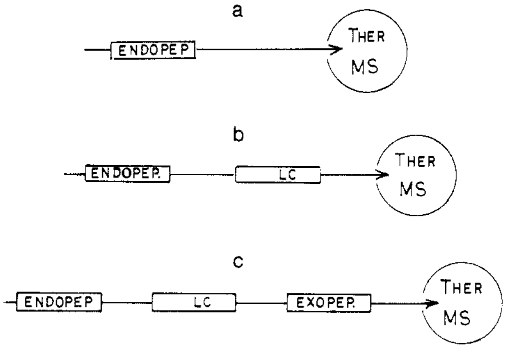


Figure 1. Block diagram of the enzyme–thermospray LC/MS sequencing instrumentation. (a) A column of immobilized endopeptidase is followed directly by the thermospray mass spectrometer. In this configuration the entire set of primary peptide fragments produced by the endopeptidase emerges as a single peak. The mass spectrum of this peak contains all of the fragment ions and can be used to ensure that all of the fragments are accounted for in later experiments. (b) The column of immobilized endopeptidase and a liquid chromatography column are attached in series, followed by the mass spectrometer. The primary peptide fragments emerge in a series of chromatographic peaks. The chromatographic map detected can also be analyzed in terms of the masses of the components present in each peak. (c) A column of immobilized exopeptidase has been added on line following the chromatographic column of (b). Each peak emerges at the same time, but the peptide fragments in that peak are then subjected to further proteolysis, from either the N or C terminus. The mass spectrum of the peak will normally contain ions of undigested primary fragment molecules plus ions of shorter (sequence) fragments from which one, two, or more amino acid residues have been removed.

The rapid analysis of protein sequences by the enzyme–thermospray LC/MS system has served as our next major goal. Basic pancreatic trypsin inhibitor (BPTI), a 58 amino acid protein, was chosen as the model for these studies. The sequence of BPTI is well established,[19] and its modest size serves as a reasonable step from the sequencing of short peptides to the analysis of large proteins. The overall design of the sequencing system used for the analysis of BPTI is similar to that described above. Moving to this larger molecule, however, introduced complexities that have required modification of each component of the system, as well as some modifications of the strategy. The most important of these are as follows.

1. Both trypsin and chymotrypsin columns have been used to generate primary fragments.

2. Columns of these two endopeptidases have also been used in concert. In such cases the second endopeptidase column is placed directly after the LC column (that is, replacing the exopeptidase column in Figure 1c). This second column generates secondary subfragments of the resolved primary fragments.

3. New LC columns have been employed. The Ultrasil column did not provide sufficient resolution of primary fragments. After a survey of other types of columns and eluent systems (to be reported elsewhere), two alternate methods have been selected: reverse-phase chromatography on a C_4 support (RP-304), and ion-exchange chromatography on poly(ethylenimine)-coated silica gel (PEI).

4. New types of immobilized exopeptidases have been introduced for the generation of complementary sequence information.

[6] Torgerson, D. F.; Skowronski, R. P.; Macfarlane, R. D. *Biochem. Biophys. Res. Commun.* **1974**, *60*, 616.

[7] Sundquist, B.; Kamensky, I.; Hakansson, P.; Kjellberg, J.; Salehpour, M.; Widdiyaskera, S.; Fohlman, J.; Peterson, P. A.; Roepstorff, P. *Biomed. Mass Spectrom.* **1984**, *11*, 242.

[8] Barber, M.; Bordoli, R. S.; Sedgwick, R. D.; Tyler, A. N. *Chem. Commun.* **1981**, 325.

[9] Vestal, M. L.; Blakely, C. R. *Anal. Chem.* **1983**, *55*, 750.

[10] Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233.

[11] Johnson, R. S.; Biemann, K. *Biochemistry* **1987**, *26*, 1209.

[12] Biemann, K.; Scobie, H. A. *Science (Washington, D.C.)* **1987**, *237*, 992.

[13] Vestal, M. L.; Fergusson, G. J. *Anal. Chem.* **1985**, *57*, 2373.

[14] Vestal, M. L. *Science (Washington, D.C.)* **1984**, *221*, 275.

[15] Pilosof, D.; Kim, H.-Y.; Dyckes, D. F.; Vestal, M. L. *Anal. Chem.* **1984**, *56*, 1236.

[16] Kim, H.-Y.; Pilosof, D.; Vestal, M. L.; Dyckes, D. F.; Kitchell, J. P.; Dvorin, H. In *Peptides Structure and Function*; Hruby, V. J., Rich, D. H., Eds.; Pierce Chemical Co.: Rockford, IL, 1983; p 719.

[17] Pilosof, D.; Kim, H.-Y.; Vestal, M. L.; Dyckes, D. F. *Biomed. Mass Spectrom.* **1984**, *8*, 403.

[18] Kim, H.-Y.; Pilosof, D.; Dyckes, D. F.; Vestal, M. L. *J. Am. Chem. Soc.* **1984**, *106*, 7304.

[19] Kassell, B.; Laskowski, M., Sr. *Biochem. Biophys. Res. Commun.* **1965**, *18*, 255.

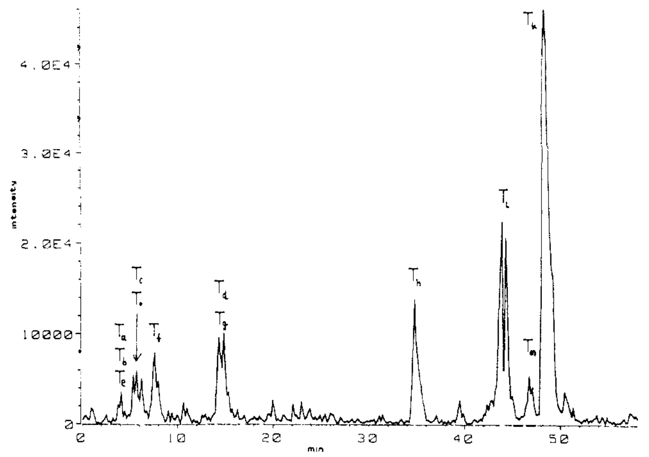


Figure 2. Total ion chromatogram of the primary tryptic fragments of CM-BPTI. The instrument is configured as in Figure 1b, using immobilized trypsin as the endopeptidase and RP-304 as the LC support. The mobile phase was 0.1 N NH_4OAc containing a gradient of 0–15% propanol. The peaks are identified by their detected components according to the listing in Table I.

Table I. Tryptic Fragments of CM-BPTI[a]

tryptic frag	mol wt	tryptic frag	mol wt
T_a	217	T_h	804
T_b	245	T_i	868
T_c	373	T_j	1186
T_d	400	T_k	1489
T_e	465	T_l	1837
T_f	521	T_m	2064
T_g	677		

[a] Calculated relationships: $T_c + T_f = T_a + T_g$, $T_m = T_b + T_l - H_2O$, $T_j = T_d + T_h - H_2O$, $T_c + T_e + T_f + T_i + T_j + T_k + T_m - 6H_2O = 6858$ (molecular weight of CM-BPTI).

Immobilized aminopeptidase M (APM) provides partial sequence data from the amino terminal end of a primary peptide fragment. N-terminal sequencing with APM may be run as a parallel experiment to C-terminal sequencing with CPY, by using an APM column in place of the CPY column. In addition, a column of immobilized carboxypeptidase B (CPB) has been prepared. This exopeptidase is used in conjunction with CPY for the generation of sequence peptides from tryptic primary fragments, which are expected to terminate in basic residues.

5. Finally, a thermospray LC/MS with a significantly larger detection range is now being used. This permits the detection and accurate identification of the larger peptide fragments generated from proteins. This instrument can be used routinely to detect ions of m/e up to 2000 amu.

Rapid sequence analysis of BPTI was carried out by using the modifications of our enzyme–thermospray LC/MS described above. The results of ten different combinations of primary fragmentation, chromatographic separation, and sequence fragment (or subfragment) generation are reported. Each experiment was carried out on 5–10 nmol of reduced, carboxymethylated, but otherwise underivatized, BPTI (CM-BPTI). Analysis of the results indicates that complete sets of tryptic and of chymotryptic fragments were detected. Every tryptic and chymotryptic fragment detected can be assigned in proper order, and exactly one-half of the inhibitor's 58 residues can be placed in correct sequence, without reference to the known structure.

Results

Analysis of the Tryptic Primary Fragments of CM-BPTI. A chromatogram of the primary tryptic fragments of CM-BPTI is shown in Figure 2. The fragments were generated by passing the protein through the immobilized-trypsin column. They were then resolved by RP-304 chromatography using ammonium acetate/propanol as the eluent. These fragments have not been further digested by an exo- or endopeptidase. The chromatogram itself represents the total ion current detected at each interval. The identity of the ion or ions constituting any given peak can be determined by displaying the actual mass spectrum summed over the appropriate interval. An example of such a spectrum is shown in Figure 3. The letters T_a through T_m above each peak in Figure 2 indicate those tryptic fragments whose ions were detected in the mass spectrum of that peak. Table I lists the tryptic fragments observed, arranged in order of increasing molecular weights, from 217 amu (T_a) to 2064 amu (T_m).[22]

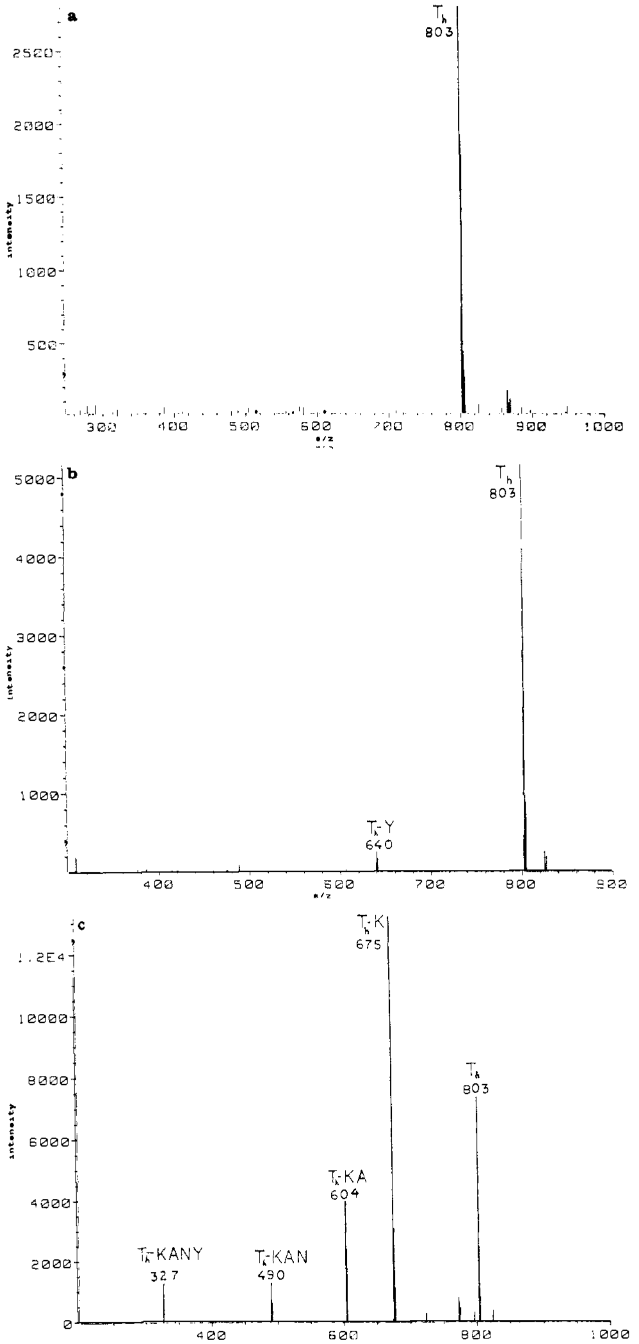


Figure 3. Mass spectra of the component emerging at 35 min in Figure 2. (a) Spectrum of the peptide without further proteolysis. The identification of the peak (T_h) corresponds to the listing of tryptic fragments in Table I. (b) Spectrum of T_h after digestion by passage through a column of immobilized APM (configured as in Figure 1c). The sequence ion peak at $m/e = 640$ is identified by the symbol for the parent primary ion minus the single-letter code of the amino acid residue corresponding to the mass difference from the parent ion. (c) Spectrum of T_h after digestion by passage through columns of immobilized CPB and CPY, in that order (configured as in Figure 1c, but with this pair of columns in the exopeptidase column position). Each successive sequence ion peak is identified by adding the single-letter code of the appropriate amino acid residue to the deleted sequence deduced in the next largest peak.

The known specificity of trypsin, of cleaving peptides on the carboxy side of lysyl and arginyl residues only, means that CM-BPTI, which possesses a total of 10 such residues, should be digested into no more than 11 discrete fragments. (The occurrence of one of these residues at the C terminus of the protein, or of a prolyl residue to the C-terminal side of any lysyl or arginyl residue, would represent additional sites not subject to cleavage and would result in fewer potential fragments.) Yet Table I lists 13 distinct fragments. Furthermore, the sum of the masses of the fragments in Table I is much larger than the molecular weight of CM-BPTI. These fragments cannot, therefore, represent a unique set corresponding to the entire sequence of the protein. Some must arise from the same region of the protein sequence and will contain duplicate sequence information. One reasonable explanation is that some of the larger fragments are a result of partial tryptic cleavages and that pairs (or higher combinations) of some of the smaller fragments should be equivalent to these larger fragments. An exhaustive search of all of the possible combinations of the fragments in Table I was carried out, comparing the masses of the combined fragments to those of the larger fragments. When such combinations are searched for, one molecule of water per junction must be subtracted. In addition to T_j (see note 22), one other probable case of partial cleavage was detected: $T_b + T_l - H_2O = T_m$.

Even when the set of tryptic fragments is corrected for the redundancies of partial cleavage, the sum of the masses of the remaining fragments is too great. A second type of duplication of information is also present. This second type of duplication arises when a peptide can give rise to two alternate, but mutually exclusive sets of fragments. In tryptic digestions, this occurs most commonly when two potential tryptic sites lie next to one another. Trypsin is not a good exopeptidase; when Arg-Arg, Arg-Lys, Lys-Arg, or Lys-Lys sequences occur in a substrate, the enzyme normally cleaves between the two basic residues, or at the C terminus of the pair, but not at both sites in a single molecule. If cleavage at each of these two alternate sites occurs at approximately the same rate, digestion of the substrate gives rise to a mixture that contains both sets of complementary fragments. These two pairs of fragments will have the same total masses. The peptides of one pair will differ from their counterparts in the other pair by the mass of an arginyl residue (156 amu) or by the mass of a lysyl residue (128 amu).[23]

A second search of the fragment masses, to find pairs of fragments that differ by 156 or 128 amu, turns up three matches. Two of these, T_a/T_c, and T_f/T_g, each differ by 156 amu; the third pair, T_b/T_c, differ by 128 amu. The peptides comprising the first two pairs can be recombined such that $T_a + T_g = T_c + T_f$; in other words, T_a,T_g and T_c,T_f can be postulated to represent the same stretch of peptide sequence. The difference of 156 amu between the corresponding fragments means that the pair of basic residues constituting the alternate cleavage sites must end in Arg (that is, the internal sequence is Arg-Arg or Lys-Arg). However, it cannot be deduced from the evidence considered thus far whether T_cT_f or T_f,T_c is the correct ordering of the sequential fragments.

There is no second pairing of peptides in the set listed in Table I which differ from one another by 128 amu. This implies that the peptides T_b and T_c, which apparently differ in composition by a lysine residue, do so by chance, and not because they are derived from the same sequence.

Table II. Partial Sequences of Tryptic Fragments of CM-BPTl from APM and CPB/CPY Digestions

tryptic frag	steps of digestion	sequence information
T_a	——	
T_b	——	
T_c	$\underline{\text{A R}}$	
T_d	$\underline{\text{I R}}$	T_a T_g $\overleftrightarrow{\text{AK}}:\overleftrightarrow{\text{R:NNFK}}$ $\underleftrightarrow{\text{AK R}}\ \underleftrightarrow{\text{NNFK}}$ T_c T_f
T_e	———	
T_f	$\underline{\text{NN K}}$	
T_g	$\underline{\text{R K}}$	
T_h	$\underline{\text{YFYNAK}}$	T_d T_h $\overleftrightarrow{\text{IIR}}:\overleftrightarrow{\text{YFYNAK}}$ $\underleftrightarrow{\text{IIR YFYNAK}}$ T_j
T_i	$\underline{\hphantom{xxxxxxxx}\text{R}}$	
T_j	$\underline{\hphantom{xxxxxxxxxx}\text{AK}}$	
T_k	$\underline{\text{AGL}\hphantom{xxxxxxxxxx}\text{R}}$	
T_l	$\underline{\hphantom{xxxxxxxxxxxxxxxxxx}\text{K}}$	$T_m = T_l + T_b$
T_m	$\underline{\hphantom{xxxxxxxxxxxxxxxxxx}\text{KAR}}$	$\langle T_b = \text{AR}\rangle$

If all of the redundancies deduced thus far, the pairs of T_b,T_l, T_d,T_f, and T_a,T_g are removed from the basis set; the sum of the remaining fragments equals the known molecular weight of CM-BPTI. This is shown in Table I.

The ability to generate a set of unique primary tryptic fragments which is also complete (in that the fragments sum to the known molecular weight of the protein) serves as strong evidence that the assignments of redundancies (or in the case of T_b and T_c, uniqueness) are correct. Confirmation of these assignments became possible once sequence data were obtained from the fragments. These data were determined by exopeptidase digestions of the individual primary fragments.

Sequencing the Primary Tryptic Fragments of BPTI. The chromatographic experiment shown in Figure 2 was repeated twice more, using the fully expanded system (Figure 1c). In one case the APM column was inserted between the RP-304 column and the thermospray detector; in the second, both the CPB and CPY columns, in that order, were placed in the same position. The total ion chromatograms in each case are essentially the same, although elution times for corresponding peaks vary slightly, depending on the number of components in the system. The mass spectra of the peaks, however, are quite different: instead of just the primary fragment ion, these spectra now contain a series of ions, corresponding to the sequence ions produced by exopeptidase digestion. A sample set of such spectra derived from the peak designated T_h in Figure 2 is shown in Figure 3.

The spectrum of the primary tryptic fragment T_h in Figure 3a shows only one significant ion, at $m/e = 803$ ($[M - H]^-$). The ion at $m/e = 825$ in each spectrum in Figure 3 represents the same peptide seen at $m/e = 803$, but with a sodium ion replacing one of the ionizable protons ($[M - 2H + Na]^-$). The spectrum in Figure 3b represents the same primary peptide fragment, now partially digested by APM. A second ion, at $m/e = 640$, can be seen. The difference of 163 amu between the primary fragment and the peptide arising from APM digestion corresponds to the mass of a tyrosyl residue. This must be the N-teriminal residue of T_h, and the smaller ion is therefore labeled T_h - Y.[24]

Figure 3c shows a mass spectrum even richer in primary sequence data. This represents the result of C-terminal digestion of T_h by the CPB/CPY column combination. The ion of the undigested primary fragment can still be seen at $m/e = 803$, but now it is not the largest signal, and a series of sequence peptide ions are clearly visible. Moving downward in mass from the primary ion T_h, these successive peaks represent the loss of 128 amu (lysyl), 71 amu (alanyl), 114 amu (asparaginyl), and 163 amu (tyrosyl), respectively. Each peak in Figure 3c has been accordingly identified as equivalent to T_h minus one or more amino

[20] *Pure Appl. Chem.* **1972**, *31*, 641.

[21] Ngo, T. T. *Biotechnology* **1986**, *4*, 134.

[22] The molecular weights listed in the tables are routinely one mass unit higher than the values of the corresponding fragment ions (normally detected as $[M - H]^{-1}$) seen in the accompanying mass spectra. One fragment not found in the chromatogram in Figure 2 is included in Table I. This is T_j, a primary fragment observed in the tryptic digest of BPTI separated on the PEI column. Fragment T_j arises from incomplete digestion between peptides T_d and T_f. The occurrence of such partial-cleavage fragments depends on the conditions of the cleavage and on the age of the trypsin column. T_j is included in Table I for completeness.

[23] Additional complications are introduced when two or more successive tryptic sites with paired basic residues or when sites with more than two consecutive basic residues occur in a protein. (Examples of both occur in histone H4.) The expected combinations can be predicted by an extension of the argument above. No such sequences occur in BPTI.

[24] The letter Y and all other letters used in sequences correspond to the IUPAC-IUB one-letter notation for amino acid sequences.[20]

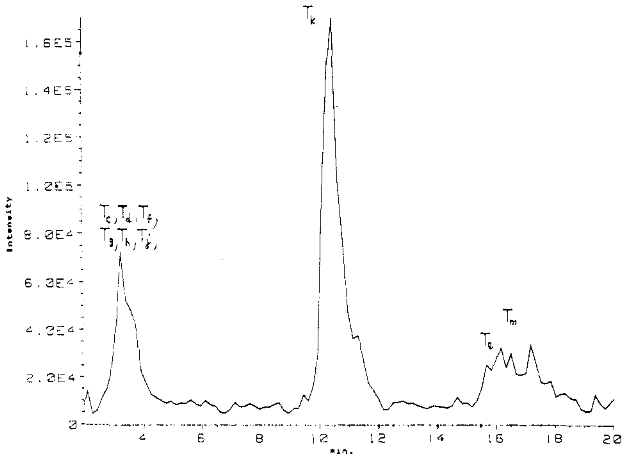


Figure 4. Total ion chromatogram of the primary tryptic fragments of CM-BPTI resolved by ion exchange on PEI. The instrument is configured as in Figure 1b. The mobile phase is a gradient of aqueous NH_4-OAc, from 0.02 M (pH 8.0) to 0.1 M (pH 7.5).

acid residues. The previous determination that the N terminus of T_h is a tyrosyl residue makes it possible to deduce the entire sequence of the peptide. The smallest of the sequence ions from CPB/CPY, T_h - KANY, is 327 amu. This fragment must have an N-terminal tyrosyl residue. Subtraction of the tyrosyl residue mass of 163 amu from the mass of T_h - KANY leaves 164 amu, the mass of the anion of the amino acid phenylalanine.[25] A phenylalanyl residue must be the second residue from the N terminus of T_h, and the sequence is Tyr-Phe-Tyr-Asn-Ala-Lys (YFYNAK).

Table II shows the entire set of results derived from digestions of the chromatographically resolved primary tryptic fragments of CM-BPTI, using APM or CPB + CPY columns. The sequences were deduced in a manner entirely analogous to those described for T_h, although in most cases the complete sequencing achieved for T_h was not possible. Some of these results merit further discussion.

The sequence data from individual peptides may be used to work out the relationships between those primary fragments which overlap one another as a result of partial cleavages. For example, it was deduced above that peptide T_m arises as a result of partial cleavage between peptides T_b and T_l. T_b is only a dipeptide and was not digested by any of the exopeptidases in any of the sequencing experiments. Digestion of T_m by the combined carboxypeptidases establishes its C-terminal sequence as -Lys-Ala-Arg. Digestion of T_l indicates lysine as the C-terminal residue. This establishes both (1) the order of the two smaller fragments, T_l–T_b, and (2) the sequence of T_b, Ala-Arg. The molecular weight of the Ala-Arg sequence exactly matches the mass of the ion detected as T_b, further confirming the correctness of the earlier hypothesis that T_m is a combination of T_l and T_b. The partial sequence data shown in Table II also confirm the assignment of T_c,T_f and T_a,T_g as complementary pairs. The mass spectra taken while the APM column was on-line indicate an alanyl residue at the N terminus of T_c and the sequence Asn-Asn- at the N terminus of T_f. The data gathered during the experiment with the CPB and CPY columns both on-line indicate arginyl at the C terminus of T_c and lysyl at the C terminus of T_f. In each case there is only one residue of the peptide not accounted for. The residual mass of the peptide T_c, once the two known residue masses are subtracted, is 141 amu (lysine anion). For T_f, the residual mass indicates phenylalanine. Applying the postulate that the paired basic residues in T_c represent the internal sites of alternate cleavage of this sequence, the order of the fragments is T_c–T_f and the sequence is Ala-Lys-Arg-Asn-Asn-Phe-Lys (AKRNNFK).

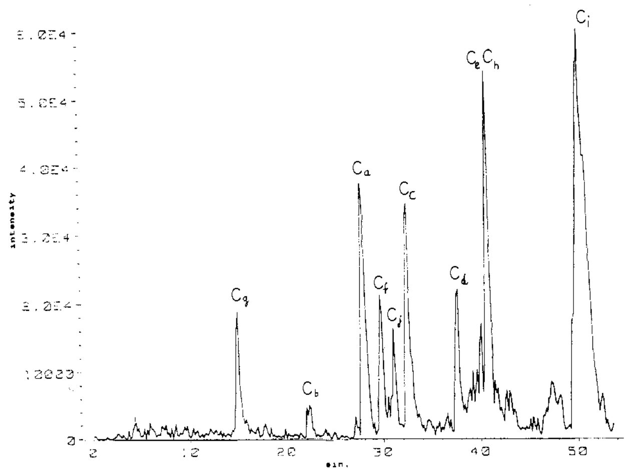


Figure 5. Total ion chromatogram of the primary chymotryptic fragments of CM-BPTI. The instrument is configured as in Figure 1b, using a column of immobilized chymotrypsin as the endopeptidase and RP-304 as the chromatographic support. A gradient of 0–15% propanol in 0.1 N NH_4OAc was used as the eluent. The peaks are identified by their detected components according to the listing in Table III.

Table III. Chymotryptic Fragments of CM-BPTI[a]

chymotryptic frag	mol wt	chymotryptic frag	mol wt
C_a	1189	C_f	1441
C_b	1179	C_g	1443
C_c	1272	C_h	1534
C_d	1371	C_i	2609
C_e	1419	C_j	2866

[a] Calculated relationships: $C_e = C_c + F - H_2O$, $C_c = C_a + Y - H_2O$ ($C_e = C_a + Y + F - 2H_2O$), $C_h = C_d + Y - H_2O$, $C_j = C_f + C_g - H_2O$, $C_e + C_i + C_j - 2H_2O = 6858$ (molecular weight of CM-BPTI).

If the arguments above are correct, it should be possible to predict the terminii of T_g: a C-terminal arginyl and an N-terminal lysyl residue. Examination of the mass spectra of the appropriate peak digests indicates precisely these results, confirming our previous interpretations.

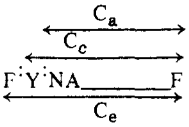
Some of the sequence data shown in Table II were actually obtained in analogous experiments using a system in which on-line digestions by APM or CPB/CPY followed PEI ion-exchange separation of the primary tryptic fragments. Although the resolution of the primary fragments in this system was not as good (see Figure 4), some exopeptidase steps proceeded better, because this system did not require the use of an organic eluent, which tends to slow enzyme digestions. One new primary fragment, T_j, was observed in this system (see footnote 22). The C-terminal sequence deduced for T_j, alanyllysine, exactly matches the C-terminal sequence of T_h, lending strong support to the inference that T_j arises from incomplete cleavage between T_d and T_h. Digestion of T_d gives two of its three residues directly. The third can be deduced from the residual mass. The complete sequence of T_h has already been derived above. Therefore the complete sequence of T_j, a nonapeptide, is determined. This is shown in Table II.

On the basis of the arguments given above, the set of tryptic fragments of BPTI can be reduced to just these six: T_e, T_i, T_j, T_k, T_m, and T_c–T_f. The ordering of these fragments within the sequence cannot be deduced without further information. The necessary information for this ordering has been determined from the chymotryptic fragments of BPTI.

Primary Chymotryptic Fragmentation of CM-BPTI. Primary fragmentation of CM-BPTI using a chymotrypsin column, followed by RP-304 chromatography and thermospray detection, gave rise to the chromatogram shown in Figure 5. The masses of the primary fragment ions detected in this chromatogram are listed in Table III. Analysis of the relative masses indicates that C_j may have resulted from incomplete digestion between C_f and C_g. No other simple combinations of this sort are apparent.

[25] Note that the mass of the entire peptide ion must equal the sum of the residue weights plus the mass of water minus one amu (the parent ions are detected as the monodeprotonated negative ions). When the masses of all but one residue are subtracted from the parent ion, the remaining mass represents an amino acid anion and is 17 amu higher than the corresponding residue mass.

Table IV. Partial Sequences of Chymotryptic Fragments of CM-BPTl from APM and CPY Digestions

chymo-tryptic frag	steps of digestion	sequence information
C_a	NA______F	
C_b	G________	
C_c	Y______F	
C_d	________VY	
C_e	F________F	
C_f	VY________	
C_g	__________A	
C_h	Y________VY	
C_i	________________Y	
C_j	________________A	$C_f + C_g = C_j$

A search through the chymotryptic primary fragments for pairs arising from alternate cleavage at paired sites can also be carried out. Chymotrypsin shows a selectivity for hydrophobic side chains in its substrates, so the most probable cleavage sites are to the C terminus of tyrosyl or phenylalanyl residues, and the search is for fragments differing by 163 or 147 amu. Three sets are found: $C_h - 163 = C_d$; $C_c - 163 = C_a$; $C_e - 147 = C_c$.

As in the case of the paired tryptic fragments T_a,T_g and T_c,T_f, described above, the pairs C_a,C_h and C_c,C_d may be tentatively assigned as complementary to one another. But here the relationship is not so clear. The fact that C_e also apparently overlaps C_c and C_a but has no complement in the observed fragment set may indicate a more complex relationship.

Removal of the apparent redundancies from the set of primary fragments in Table III does not reduce it to a unique sequence set. The sum of the fragment masses still exceeds the known molecular weight of CM-BPTI. This is true whether C_a,C_h and C_c,C_d are considered to be complementary pairs or not. The set does simplify enough, however, to make a reasonable assignment of a probable unique set. If C_a and C_c are taken to be subfragments of C_e, if C_d is assigned as a subfragment of C_h, and if the pair C_f,C_g is assumed to be combined in C_j, then the remaining peptides are C_b, C_e, C_h, C_i, and C_j. A set composed of three of these, C_e, C_i, and C_j, has a weight (the three peptide masses − $2H_2O$) precisely equal to that of the protein. These can be tentatively assigned as a complete set and then tested by comparing peptide sequence data and by tryptic subfragmentation of the entire array of chymotryptic primary fragments.

Table IV shows the results of partial sequencing of the primary chymotryptic fragments by inserting either an APM or a CPY column into the system following the RP-304 column (Figure 1c). Analyses of the spectra of each chromatographic peak are carried out just as described above for the exopeptidase digestions of the tryptic fragments. From the sequences in Table IV, some of the calculated relationships between the primary fragments can be confirmed.

C_a, C_c, and C_e were assigned as successively larger fragments arising from the same protein segment. All possess identical C-terminal phenylalanyl residues. C_c and C_e are extended at the N terminus by tyrosyl and phenylalanyl residues, respectively, exactly the results predicted from the relative masses of C_a, C_c, and C_e (Table III). If these two residues, Phe-Tyr-, are combined with the Asn-Ala- dipeptide sequence found at the N terminus of C_a, it is possible to construct a tetrapeptide sequence for the N terminus of C_e: Phe-Tyr-Asn-Ala-. This same tetrapeptide sequence appears in one of the fully sequenced tryptic fragments, T_j, thus making a very strong case in favor of the supposed relationship of the three chymotryptic fragments.

If the assignments just discussed are true, then C_a,C_h and C_c,C_d do not represent a complementary pair. C_h and C_d have identical C-terminal sequences and differ from one another by the N-terminal tyrosyl residue by which C_h is larger. But C_a and C_c also differ at their N terminii and so cannot arise from an overlap with the other fragment pair.

One other important set of relationships can be confirmed from the data in Table IV. C_g has an alanyl residue at its C terminus. On the basis of the known preference of chymotrypsin for cleavage to the C-terminal side of residues with bulky hydrophobic side chains, it is unlikely that this alanyl residue was a cleavage site. This implies that C_g is the C-terminal chymotryptic fragment of the protein, and the C-terminal subfragment of C_j as well.

Table V. Tryptic Subfragments of the Chymotryptic Fragments of CM-BPTl

chymotryptic frag	mol wt	tryptic subfragments		
		no.	MW	identity
C_a	1109	1	331	C_d
		2	796	$C_{c2} = C_{e2}$
C_b	1179	1	373	$C_{f1} = C_{j1} = T_c$
		2	393	C_{f2}
		3.	449	
C_c	1272	1	494	C_{h1}
		2	796	$C_{a2} = C_{e2}$
C_d	1371	1	331	C_{a1}
		2	1058	C_{h2}
C_e	1419	1	641	
		2	796	$C_{a2} = C_{c2}$
C_f	1441	1	373	$C_{b1} = C_{j1} = T_c$
		2	393	C_{b2}
		3	711	C_{j4}
C_g	1443	1	465	C_{j2} = Te
		2	996	
C_h	1534	1	494	C_{c1}
		2	1058	C_{d2}
C_i	2609	1	245	T_b
		2	563	
		3	1837	T_l
		4	2064	T_m
C_j	2866	1	373	$C_{b1} = C_{f1} = T_c$
		2	465	$C_{g1} = T_e$
		3	521	T_f
		4	711	C_{f3}
		5	868	T_i

Tryptic Fragmentation of the Chymotryptic Primary Fragments. Given a unique set of primary tryptic fragments and a proposed unique set of primary chymotryptic fragments, each with partial sequence data determined, the proper alignment of these fragment sets requires overlaps between the two sets to be established. Only one overlap has been proposed on the basis of the sequence data alone, between fragments T_j and C_e. If this overlap is accepted, then of the three chymotryptic fragments, C_e, C_i, and C_j, proposed as constituting a complete unique set, C_e cannot be the N-terminal fragment. If C_j, by virtue of its C-terminal alanyl residue, is assigned as the C-terminal fragment, then the ordering C_i-C_e-C_j naturally follows. The tyrosyl residue at the C terminus of C_i is also consistent with the proposed overlap sequence in the tryptic peptide T_h.

An examination of the tryptic subfragments of the primary chymotryptic fragments both confirms this alignment and permits unambiguous ordering of all of the observed primary tryptic fragments. The tryptic subfragments are generated by replacing the exopeptidase column in the chymotrypsin primary fragment analysis system (Figure 1c) with a trypsin column. The total ion chromatogram is virtually identical with that of the primary chymotryptic fragments in Figure 5. However, this time the mass spectrum of each peak contains tryptic subfragments of each primary fragment, instead of sequence ions. These are listed in Table V.

Each primary chymotryptic fragment was cleaved into at least two tryptic subfragments. In those cases where only two tryptic subfragments are generated, indicating a single internal tryptic cleavage site, neither subfragment can normally be expected to match a primary tryptic fragment. Table V shows one exception to this rule, and it is significant. C_g is cleaved into only two tryptic subfragments, but one of them, C_{g1}, matches T_e. Both C_g and T_e must have one terminus in common, but that terminus cannot have been the result of both chymotryptic and tryptic cleavage at the same site. Their common terminus must have been one of the protein's terminii as well.

This conclusion is fully consistent with the earlier identification of C_g as the C-terminal primary chymotryptic fragment. T_e is the C-terminal primary tryptic fragment. Examination of the mass

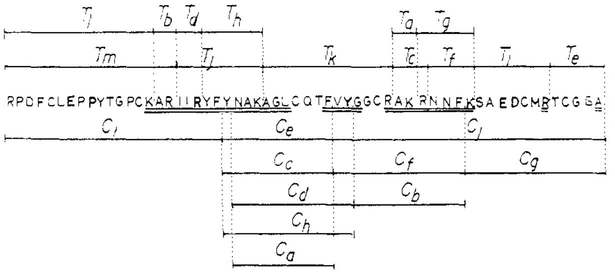


Figure 6. An alignment of the tryptic and chymotryptic primary fragments detected in this study against the single-letter-code amino acid sequence of BPTI. The labels identifying each primary fragment correspond to the listings in Tables I and III. The residues that are underlined correspond to those which could be placed correctly as a result of these experiments.

spectrum of the tryptic subfragments of C_j also shows the presence of an ion identical in mass with T_e (C_{j2}). This observation confirms the independent evidence discussed earlier, identifying C_g as a chymotryptic subfragment of C_j.

The tryptic subfragments of C_j deserve closer scrutiny. The five listed in Table V account exactly for the total mass of C_j. Consistent with its position at the protein's C terminus, all but one (C_{j4}) of the tryptic subfragments of C_j corresponds to a primary tryptic subfragment (Table V). Two subfragments, C_{j1} and C_{j4}, also correspond to two of the tryptic subfragments (C_{f1} and C_{f3}, respectively) of the primary chymotryptic fragment C_f. This confirms the identity of C_f as a chymotryptic subfragment of C_j. But the tryptic subfragment C_{j4} (which corresponds to C_{f3}) is not a primary tryptic fragment. It must come from the N terminus of C_j (and of C_f). This is fully consistent with the assignment of C_f at the N terminus of C_j.

Subfragment C_{j2} (T_e) has already been placed at the protein's C terminus. Two tryptic subfragments of C_j, T_f (C_{j3}) and T_i (C_{j5}), remain to be assigned. One should represent the overlap segment between C_f and C_g and would not be expected to be found among their tryptic subfragments. But neither is seen among them. This requires a more detailed analysis.

There are two tryptic subfragments of C_f and C_g that do not correspond to tryptic subfragments found in C_j: C_{f2} and C_{g2}. The mass of T_i (868 amu) is very close to that of C_{g2} (996 amu), which has already been identified as the N-terminal tryptic subfragment of C_g. The difference between T_i and C_{g2}, 128 amu, corresponds exactly to one lysyl residue. This lysyl residue would logically be assigned to the N terminus of C_{g2}, where it would not be a tryptic site. Thus the release of T_i from C_g would not be expected. If this placement of T_i is correct, then (1) T_f is the overlap peptide between C_f and C_g and must have a C-terminal lysyl residue (2) the unassigned tryptic subfragment of C_f, C_{f2}, must correspond to T_f minus the lysyl residue. Both predictions are confirmed by the data.

All of the tryptic subfragments of C_j can be arranged in order. The placement of the four which are also primary tryptic fragments is shown in Figure 6. The position of the primary chymotryptic fragment C_b, which overlaps this region and shares common tryptic subfragments with C_f and C_j, is also shown in Figure 6. The arguments for the placement of C_b will not be presented here, but they are entirely analogous. They also serve to further confirm the arrangement of subfragments in C_j.

The primary tryptic fragment that bridges C_e and C_j is the only overlap fragment not yet identified. The basis set of primary tryptic fragments has already been simplified to a unique sequence set: T_e, T_i, T_j, T_k, T_m, and T_c–T_f. Of these, all but T_j and T_k have been identified as tryptic subfragments in either C_i or C_j. On the basis of common sequence, T_j has already been proposed to be the C_i,C_e overlap fragment. If the proposed ordering of the three unique chymotryptic fragments is correct and if C_{j4} is indeed the N-terminal subfragment of C_j, then the mass of C_{j4} plus the mass of one of the two tryptic subfragments of C_e must equal mass of T_k (plus H_2O). This is indeed the case: $C_{j4} + C_{e2} = T_k + H_2O$. To round off this assignment, C_{e1}, the remaining tryptic subfragment of C_e, can be matched with C_{i2} (the only tryptic subfragment of C_i which is not also a primary tryptic fragment) to match the mass of T_j, already identified as the overlap fragment on the basis of its sequence. The evidence for the alignment C_i-C_e-C_j is now overwhelming. Placement of the remaining tryptic and chymotryptic fragments is straightforward. The alignment of all fragments and the placement of all sequenced residues are shown in Figure 6.

Discussion

The generation of sequence data by the enzyme–thermospray LC/MS method of analysis is rapid. In only a few experiments, none of which required more than an hour of instrument time, and some of which were complete in less than 5 min, it has been possible to gain enough information to deduce the relationships of all of the tryptic and chymotryptic peptides of CM-BPTI and to place exactly half of the amino acids in their proper sequential order.

Material consumption using this method is modest. The total protein required for the ten experiments reported was under 100 nmol (5–10 nmol per injection; in the case of BPTI this is less than 600 μg). Furthermore, all of the experiments were run on protein samples that were not modified beyond the reduction and carboxymethylation of the disulfide bonds.

Several other aspects of this method merit special mention:

(1) Each experiment generates sequence information from a large number of sites in the protein, internal as well as terminal. Application of the method is not blocked by terminal modifications such as N-acetylation or C-amidation.

(2) Although only trypsin and chymotrypsin were used for this study, there is a wide arsenal of endopeptidases available for primary fragment production. Thus many alternatives exist for the targeting of specific sites within a protein for direct sequence analysis.

(3) Each primary fragment can be sequenced from both the N and C terminus. This increases the absolute amount of sequence data generated. It also significantly improves the possibility of identifying fragment overlaps.

(4) Digestions of the protein substrate by the immobilized enzymes used in this study were remarkably efficient. Very few peptides arising from partial cleavages were observed, indicating that the tryptic and chymotryptic cleavages were virtually complete in the time required for the protein to pass through the endopeptidase column. This time is approximately 1 min, judged by the width of the peak emerging from the column. Standard protocols for the digestion of proteins by endopeptidases normally call for incubation times of a few hours.

A detailed explanation of the logic used in analyzing the mass spectral data was presented in the results section. Virtually everything explained there is capable of being translated into relatively simple algorithms. Direct analysis of the experimental data and generation of sequence information by computers are therefore entirely reasonable. Such analyses could even be carried out in an on-line procedure. This would permit the current status of the sequence information to be displayed and could lead to suggestions of additional or alternate experiments by either the experimenter or an instrumental expert system while the analysis was in progress. A large proportion of the experimental protocols are also clearly amenable to automation.

In its present form the enzyme–thermospray LC/MS method of sequence analysis is still subject to some limitations and cannot be expected to work in all cases. Very large proteins may be too complicated for analysis without prefragmentation. Extremely hydrophobic proteins, membrane proteins for example, may require special techniques to render them suitable for analysis in an essentially aqueous system with volatile buffer components. Some core protein sequences may simply prove intractable to any combination of enzymes currently known.

The arguments used in this study to deduce the complete peptide fragment set required one extrinsic bit of information, a reasonably accurate molecular weight. Such weights are not within the range of the instrument used in these studies, although might be determined on magnetic sector instruments using methods already reported. The weight of BPTI is well established,[19] but in those cases where the molecular weight of the protein under study is not accurately known, there may well be a greater degree of ambiguity in the deduction and ordering of a complete fragment set.

The most important problem for de novo sequencing, however, is the fact that it is not possible to differentiate between leucine and isoleucine from the masses of their residues. Since no fragmentations of the amino acids themselves are carried out under the current methods, the identification of a residue of mass 113 amu involves a twofold ambiguity. In the sequencing study reported here, the correct residue identities were listed for convenience, but each should properly read Leu/Ile (L/I). A similar equality in mass (at least at the unit resolution level) between lysine and glutamine should also be noted. The enzymological differences in the behavior of these residues (especially toward trypsin) has permitted us to differentiate these two unambiguously thus far, but they could possibly give rise to confusing results in future cases. Here chemical derivatization of the side chains may prove useful. Better yet, the use of high-resolution mass spectrometry should solve such problems.

Despite these current limitations, enzyme–thermospray LC/MS analysis of proteins has a wide variety of potential applications immediately available. This paper has documented its usefulness in rapid partial protein sequence analysis, where it should be able to complement effectively those sequencing methods that are already well established.[1,3-5]

Even the availability of powerful methods of total protein sequence analysis through recombinant DNA techniques does not eliminate the need for rapid partial sequencing of the proteins themselves. The method presented here can positively complement gene sequencing in several ways: It can be used to identify stretches of sequence in a novel protein which could serve as the basis for the synthesis of the DNA probes needed for the identification and isolation of the gene for total sequencing; it can be used to confirm deduced placements of, and changes in the reading frame associated with, intervening sequences in genes; and in the case of a gene whose sequence has been determined, but whose protein product and the conditions of its expression are unknown, this method can be used to scan rapidly through a number of protein candidates for matches to the predicted sequence.

In addition to sequence analysis, the enzyme–thermospray LC/MS method should prove highly versatile as a means for peptide and protein mapping. As was demonstrated in this study, the maps produced are multidimensional: they possess data in the form of elution patterns, primary fragment masses, and fragment sequence data. They should prove especially valuable in those cases where the sequence of a molecule is known, or thought to be known. The location of disulfide bridges within a protein, the identification of which protein variant or isozyme is being expressed by a particular individual or at a particular stage of development, and the identification of posttranslational modifications in a protein may all be carried out by mapping techniques available through this method.

Mapping is also often used in the analysis of synthetic products. With the availability of rapid, multidimensional mapping, the analysis of peptides and proteins prepared by chemical synthesis, to ensure against the presence of amino acid deletions, insertions, or modifications, can become routine. Similarly, the protein products of genetic engineering can be analyzed to ensure that the molecule synthesized has the structure anticipated. The type of confirmatory information afforded by this technique should prove especially valuable in the characterization of complex synthetic biomolecules intended for therapeutic applications.

Experimental Section

Endopeptidase and Exopeptidase Columns. Trypsin, chymotrypsin, CPB, and APM were purchased from Sigma Chemicals (St. Louis, MO). A highly purified sample of CPY was the gift of Carlsberg Biotechnology Ltd. (Copenhagen, Denmark). Columns containing these enzymes in an immobilized form were prepared in our laboratories by procedures that will be described in detail elsewhere. Briefly, trypsin and chymotrypsin were immobilized on *N*-hydroxysuccinimide glycophase controlled pore glass beads (NHS/GCPG) by the manufacturer's protocols. APM, CPB, and CPY were attached to GCPG activated by 2-fluoro-1-methylpyridinium toluenesulfonate by an adaptation of the method of Ngo.[21] GCPG and NHS/GCPG were purchased from Pierce Chemical Co. (Rockford, IL). Approximate loading levels of the enzymes were 20–50 μmol/g. The beads bearing immobilized enzymes were packed in stainless steel tubes (0.2 × 10–20 cm) by the slurry method under pressures of up to 4000 psi.

Liquid Chromatography. LC separations of primary tryptic fragments were performed on a Hi-Pore RP-304 (C-4 reverse phase, 5 μ, 4.6 mm × 20 cm) column purchased from Bio-Rad Laboratories (Richmond, CA) or on poly(ethylenimine)-coated silica gel ion-exchange columns custom prepared for Dr. William Hutchens at the Reproductive Research Laboratory, Baylor College of Medicine, by J. T. Baker Research Products (Phillipsburg, NJ).

The separation of CM-BPTI tryptic fragments on RP-304 was accomplished by eluting with a 60-min linear gradient of 0–15% propanol in 0.1 N NH_4OAc, pH 6.9, at a flow rate of 1 mL/min. For the separation of tryptic fragments on PEI, a 20-min linear gradient from 20% to 50% of 0.1 N NH_4OAc, pH 7.5, in 0.02 N NH_4OAc, pH 8.0, was used at a 1 mL/min flow rate.

The separation of the primary chymotryptic fragments of CM-BPTI was performed on the RP-304 column using 0.1 N NH_4OAc, pH 6.9 (A), and 15% propanol in 0.1 N NH_4OAc, pH 6.9 (B), at a flow rate of 1 mL/min. An initial gradient from 0% to 5% B in A was followed by a 35-min gradient from 5% to 100% B. Elution with pure B was continued for an additional 20 min.

Preparation of Carboxymethylated Basic Pancreatic Trypsin Inhibitor. The BPTI used in these studies was a homogeneous protein preparation (Trasylol), purified from natural sources and donated by Bayer, A. G., Wuppertal, F. R. G. It was reduced at 37 °C in 6 M guanidine·HCl/0.2 M Tris·HCl, pH 8.5, by the addition of a 25 M excess of dithiothreitol (Pierce Chemical Co.). After 2.5 h of reaction all sulfhydryl groups were blocked by the addition of a 110% molar excess (over dithiothreitol) of iodoacetate at pH 8.5, to form S-carboxymethylated derivatives. Reduced and carboxymethylated BPTI (CM-BPTI) was isolated by gel filtration on a column of Sephadex G-50 (fine) equilibrated to and developed with 2% HOAc. The lyophilized solid recovered was dissolved in a minimum of 20% acetic acid (370 μL/μmol CM-BPTI) and diluted with deionized water to a protein concentration of 5×10^{-4} M. Injections of 10–20 μL of this solution were used in the sequencing experiments.

The thermospray LC/MS instrument used in these studies was a Hewlett-Packard quadrupole mass spectrometer (Model 5988A) equipped with a Vestec Thermospray LC/MS Interface, specially modified for peptide work on this instrument (Vestec, Houston, TX). The primary modifications were the addition of a second block heater, adjacent to the original one, and a slight enlargement of the MS source slit. All spectra reported were obtained in the negative ion mode, which gave better response and less background than the positive ion mode. The mobile phases were delivered by a HPLC gradient system from Scientific Systems, Inc. (State College, PA).

Acknowledgment. We thank the National Institutes of Health (Grant GM-32772) for support of this work. We also thank Dr. Lilly Yang for her expert help in the early stages of this project and Dr. William Hutchens and Mr. Chee-Ming Li for supplying the PEI column used in this work. The generous gifts of highly purified samples of CPY from Carlsberg Biotechnology and of BPTI (Trasylol) from Bayer A.G. are gratefully acknowledged.

Registry No. BPTI, 9087-70-1; APM, 9054-63-1; CPB, 9025-24-5; CPY, 9046-67-7; endopeptidase, 9001-92-7; exopeptidase, 9031-96-3; trypsin, 9002-07-7; chymotrypsin, 9004-07-3.