

An algorithm to determine protein sequence alignment by utilizing data obtained from a peptide mixture and individual peptides

Carlo Caporale,^{1*} Ciro Sepe,¹ Carla Caruso,¹ Pasquale Petrilli² and Vincenzo Buonocore¹

Abstract

With the aim of limiting peptide purification steps and unambiguously ascertaining protein sequences, we have designed and implemented on a personal computer an algorithm to determine sequence alignment by utilizing data obtained from automatic Edman degradation performed on a single peptide mixture and individual peptides. The protein under study is digested by two different hydrolysis methods and fragments are just isolated from one mixture and sequenced, while the second mixture is submitted unfractionated to sequence analysis. The algorithm provides for the exact alignment of the individual peptides using the mixture data for the overlapping. We report an example of application of this approach by utilizing experimental data obtained from a protein of known sequence.

Introduction

An approach to limit peptide purification steps in assessing protein sequence was first proposed by Gray in 1968, who designed an algorithm working on phenylthiohydantoin (PTH) amino acid data deriving from sequence analysis of unfractionated peptide mixtures obtained digesting the protein by various hydrolysis methods (Gray, 1968). This algorithm, consisting of determining the residue(s) common to each consecutive step of Edman degradation in all sets of data after their correct alignment on the basis of the specificity of hydrolytic agents, suggested diverse approaches using data deriving from unfractionated protein digests (Fairwell *et al.*, 1970; Cannon and Lovins, 1972; Biemann, 1980; Matsuo *et al.*, 1981; Kitagishy *et al.*, 1981; Shimonishi *et al.*, 1981; Herlihy and Biemann, 1981; Kitagishy *et al.*, 1982; Erickson and Jardine, 1986). These procedures, revised by Petrilli and Colosimo (Petrilli and Colosimo, 1990), are mainly based on the use of mass spectrometry data, owing to the feature of this technique of working on peptide mixtures. Furthermore, algorithms to determine sequence alignment utilizing Edman degra-

tion data of individual peptides combined with amino acid composition data (Petrilli, 1985) or fast atom bombardment mass spectrometry (FAB-MS) data from peptide mixtures (Petrilli *et al.*, 1991) have been reported.

Recently, we proposed an algorithm based on the original Gray's idea not requiring FAB-MS data (Caporale *et al.*, 1993) and applied it in assessing the sequence of a trypsin inhibitor from wheat kernel coded WTI (Poerio *et al.*, 1989; Poerio *et al.*, 1994). We showed that the interpretation of the sequence data obtained from peptide mixtures deriving from the hydrolysis of the pyridylethylated protein (PyEt-WTI) by CNBr, endoprotease Lys-C and endoprotease Asp-N was very simple. The complete sequence of the protein was determined utilizing the above three sets of data and just one uncertainty was found at one position (Cys or Arg). The theoretically derived FAB-MS masses of all plausible peptides present in the mixtures and matching the output sequence were also furnished to the user in order to clarify doubts performing a FAB-MS analysis. In fact, the possible simultaneous overlap of more than one residue in the aligned sets of data generates uncertainties. However, these uncertainties can be also resolved by this algorithm supplying additional sequence data deriving from further hydrolysis method(s) (Caporale *et al.*, 1993).

A different approach can be used to exclude all possibility of doubts. In fact, it is possible to reconstruct the entire amino acid sequence of a protein from sequencing data derived from a complete set of individual peptides together with the sequencing information from a mixture derived from another hydrolysis method. In this paper, we describe an algorithm implemented on a personal computer which is based on this idea allowing one to assess protein sequences unequivocally. Obviously, no doubt is possible when the primary structure of individual peptides is determined and data from a peptide mixture are utilized for the alignment in order to limit purification steps and sequence analyses. This algorithm represents an alternative to the previous one (Caporale *et al.*, 1993) when FAB-MS data or numerous hydrolytic agents are not available. We also furnish an example of application using experimental data obtained from individual peptides and a mixture deriving from different digestions of PyEt-WTI.

¹Dipartimento di Agrobiologia ed Agrochimica, Università della Tuscia, via S. Camillo de Lellis, 01100, Viterbo, Italy and ²Istituto di Industrie Agrarie, Università di Napoli, Portici, Italy

*To whom reprints requests should be sent

System and methods

Triticum aestivum, pure variety San Pastore, was kindly supplied from Istituto Nazionale per la Cerealcoltura (Rome, Italy). Pulsed liquid-phase automatic sequencer (model 477A) equipped with on-line PTH analyser (model 120A) and relative reagents were from Perkin-Elmer–Applied Biosystems. Sequencing grade protease V8 from *Staphylococcus aureus* was from Boehringer-Mannheim Italia SpA. HPLC procedures were carried out on a Beckman GOLD apparatus equipped with a variable-wavelength monitor model 166. All other reagents were of analytical grade.

Purification, reduction and S-alkylation of inhibitor WTI

Purification of inhibitor WTI was obtained as previously reported (Poerio *et al.*, 1989). Reduction and S-alkylation of the protein by 4-vinylpyridine was performed as previously described (Caporale *et al.*, 1993; Poerio *et al.*, 1994).

Peptide mixture

Experimental procedures concerning chemical cleavage of the pyridylethylated protein by CNBr as well as details on PTH-amino acids pmole data achieved at each step of the sequence analysis performed on the obtained mixture were previously reported (Caporale *et al.*, 1993).

Individual peptides

Experimental procedures concerning PyEt-WTI digestion by protease V8 as well as details on isolation and sequence analyses of the individual peptides were previously described (Poerio *et al.*, 1994).

Hardware and software

Programs were written using Microsoft QuickBASIC

(version 1.00b) and implemented on an Apple Macintosh Classic computer. The operative system was System 6.7. No problem was observed in running the compiled application using System 7.1.

Algorithm

With the advent of the modern pulsed-liquid phase sequencers equipped on-line with phenylthiohydantoin amino acid derivative analysers, it is possible to identify more than one amino acid residue at each step of the Edman degradation. Because the chemistry for Edman degradation has been highly optimised, it is now possible to adopt the sequencing of mixtures of peptides as a rapid, general strategy to resolve specific problems very quickly. In fact, owing to the possible simultaneous presence of the same residue(s) at various sequencing steps and to the presence of short peptides which are completely degraded in few steps, the interpretation of the sequence data is simplified. Of course, cleavage methods producing large fragments should be preferentially chosen with the aim of further limiting sequence analyses of individual peptides; moreover, sequence data from a mixture containing few large fragments are more easily interpreted. This allows one to construct appropriate algorithms for optimizing the performance of the instrument.

In Figure 1 are shown one-letter code amino acid sequences of the three individual peptides isolated from PyEt-WTI protease V8 mixture (Figure 1A) determined as already described (Poerio *et al.*, 1994) as well as all PTH-amino acids identified on the basis of pmole data at each of the 28 steps of the Edman degradation performed on PyEt-WTI CNBr mixture (Figure 1B) as previously reported (Caporale *et al.*, 1993). They represent the data necessary to reconstruct the sequence of the protein.

A)																												
a) N F C K R R C T P A R																												
b) E E A M P S A W P C C D E																												
c) C G T C T R M I P P R C T C M D V S P S G C H P A C K N C V Q T T L G G R D V F W C M L R I E																												
B)																												
	Edman step																											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
D	S	S	T	N	T	C	D	E	A	G	T	N	T	R	Q	T	T	L	G	G	R	D	V	F	W	C	T	
E	E	A	E	S	G		T	R	R	C	K	P	A	V	T													
P	R	P	R	P	C		H	P	C		C	C																
I	P	I	P	C	F		K																					
L	V		W																									

Fig. 1. Amino acid sequence of PyEt-WTI individual peptides obtained by protease V8 hydrolysis (A) and PTH amino acids identified at each sequencing step performed on PyEt-WTI peptide mixture obtained by CNBr treatment (B). Cys residues were identified as PTH-PyEt-Cys.

Implementation and algorithm description

Input/Edit routine

The data necessary to determine protein sequence are inputted from the keyboard and stored on the disk. In particular, the user must input the determined sequences of individual peptides obtained by the first hydrolysis method as well as all the amino acids identified at each step of Edman degradation performed on the mixture deriving from the second hydrolysis method. Furthermore, the user must furnish some information about the digesting agents used for both hydrolysis methods; in particular, the expected amino acids present at sites of hydrolysis on the basis of the specificity of the hydrolytic agent should be indicated (e.g. E for the first hydrolysis method if protease V8 was used in producing individual peptides and M for the second hydrolysis method if CNBr was used in yielding peptides mixture) together with their position

(N- or C-terminal) in the produced peptides (e.g. C-terminal in the above examples). Finally, it should be specified if the PTH derivatives of the amino acid(s) present at hydrolysis sites are identifiable in the sequence analysis of the mixture (e.g. PTH-Met is not identifiable in the mixture deriving from CNBr treatment, owing to the formation of homoserine and homoserine lactone; the retention time of PTH-homoserine is very similar to that of PTH-Thr in the HPLC system of analysis of PTH-amino acids).

Searching section

The algorithm we designed to align individual peptides for assessing the sequence of the protein can be summarized as shown in Figure 2. First of all, the algorithm establishes if individual peptides from protease V8 digestion contain potential sites of hydrolysis specific for the hydrolytic agent used to obtain the unfractionated CNBr mixture. In

Individual V8 peptides	Possible searching ways due to cuts at C-Terminal site M	Start-End pairs of Edman steps in CNBr mixture
a) NFCKRRCTPAR Possible CNBr cut ↓	NFCKRRCTPAR	(no site) 5 - 15
b) EEAMPSAWPCDE Possible CNBr cut ↓	EEAMPSAWPCDE EEAM/ PSAWPCDE	(no cut) not found (one cut) 1 - 4 1 - 9
c) CGTCTRMIPPRCTCMDVSPSGCHPACKNCVQTLGGRDVFWMRLIE Possible CNBr cut ↓ Possible CNBr cut ↓ Possible CNBr cut ↓	CGTCTRMIPPRCTCMDVSPSGCHPACKNCVQTLGGRDVFWMRLIE CGTCTRM/ IPPRCTCMDVSPSGCHPACKNCVQTLGGRDVFWMRLIE CGTCTRMIPPRCTCM/ DVSPSGCHPACKNCVQTLGGRDVFWMRLIE CGTCTRMIPPRCTCMDVSPSGCHPACKNCVQTLGGRDVFWM/ LRIE CGTCTRM/ IPPRCTCM/ DVSPSGCHPACKNCVQTLGGRDVFWMRLIE CGTCTRM/ IPPRCTCMDVSPSGCHPACKNCVQTLGGRDVFWM/ LRIE CGTCTRMIPPRCTCM/ DVSPSGCHPACKNCVQTLGGRDVFWM/ LRIE CGTCTRM/ IPPRCTCM/ DVSPSGCHPACKNCVQTLGGRDVFWM/ LRIE	(no cut) not found (one cut) 10 - 16 not found (one cut) not found not found (one cut) not found 1 - 4 (two cuts) 10 - 16 1 - 8 not found (two cuts) 10 - 16 not found 1 - 4 (two cuts) not found 1 - 28 1 - 4 (three cuts) 10 - 16 1 - 8 1 - 28 1 - 4

Peptides alignment:

Peptide b (truncated way 1-4/1-9; start-end pair 1-9)-Peptide c (truncated way 10-16/1-8/1-28/1-4; start-end pair 10-4)-Peptide a (continuous way 5-15; start-end pair 5-15)

Protein sequence:

1 5 10 15 20 25 30 35 40 45 50 55 60 65 70
EEAMPSAWPCDECGTCTRMIPPRCTCMDVSPSGCHPACKNCVQTLGGRDVFWMRLIENFCKRRCTPAR

Fig. 2. Elucidation of the searching algorithm: sequence data from individual V8 protease peptides and CNBr mixture were utilized. Numbers in bold face indicate the defined start-end Edman step pairs of the complete sequence of individual peptides in the mixture data.

Edman step (mixture)																												
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
D	S	S	(T)	N	T	C	D	E	A	G	T	N	T	R	Q	T	T	L	G	G	R	D	V	F	W	C	(T)	
E	E	A	E	S	G	(T)	R	R	C	K	P	A	V	(T)													(M)	
P	R	P	R	P	C	H	P	C			C	C		(M)														
I	P	I	P	C	F	K																						
L	V		W		(M)		(M)																					
Individual peptides																												
E	E	A	M/																									peptide b
P	S	A	W	P	C	C	D	E																				peptide c
I	P	P	R	C	T	C	M/																					
D	V	S	P	S	G	C	H	P	A	C	K	N	C	V	Q	T	T	L	G	G	R	D	V	F	W	C	M/	
L	R	I	E																									peptide a

Fig. 3. Correct alignment of the sequence of individual V8 peptides in the CNBr mixture data. M residues not identifiable in the mixture data are indicated in parenthesis. They replace T residues in parenthesis representing homoserine.

the example shown in Figure 2, V8 peptide 'a' does not contain any site (M) recognized by CNBr treatment, while V8 peptides 'b' and 'c' contain one and three sites, respectively. When the position of such sites has been identified, the search of the sequence of the individual peptides in the sequence data of the mixture will start. It should be noted that, as expected, M residues were not present in the mixture data (Figure 1B) owing to the formation of homoserine. Consequently, M residues present in individual peptides could not be found at any Edman step. For this reason, the algorithm assumes that methionine can be present at all steps of the mixture data. Now, the algorithm will analyze all the possibilities of hydrolyzing the individual peptides at the definite sites specific for CNBr treatment, taking into account that partial or no cleavage could have been occurred on the molecules present in the mixture (Figure 2). Of course, this analysis is necessary as the algorithm must not be dependent on the effectiveness of the hydrolytic agents and the sequence of the individual peptides must be searched in the data of the unfractionated mixture in all possible ways. The sequence of peptide 'a' not containing any CNBr sites will be searched just by one continuous way starting from the first Edman step (i.e. this 11 residues sequence is searched in CNBr mixture data from step 1 to 11, from 2 to 12 and so on), while the sequence of peptides 'b' and 'c' will be searched both by continuous (no cut) and 'truncated' ways corresponding to all possible cleavages on the molecules present in the mixture. Searching by truncated ways implies that, when all residues preceding an hydrolysis site (in Figure 2) have been found by a continuous way starting from any step, next residues of individual peptide sequence will be searched restarting from the first step in mixture data, since, if cleavage occurred, the examination should continue from the N-terminal residue of another fragment present in the mixture.

Result output furnishes the start-end pair(s) of Edman

steps defined for each individual peptide when its complete sequence has been found in the mixture data following the right searching way(s). Of course, more than one pair could be identified for each peptide if partial cleavages occurred on molecules present in the mixture and the corresponding PTH data were inputted in more than one way. However, the alignment of individual peptides results obvious in any case looking at the consecutive start-end pairs and their searching ways. Just one pair was found for each V8 peptide. The analysis by continuous and truncated ways shown in Figure 2 indicates that complete cleavages of the protein occurred at M sites by CNBr treatment (complete cut at the only site of peptide 'b' and complete cuts at the three sites of peptide 'c'). Peptides alignment was immediate: peptide 'b', starting from the first Edman step, was the N-terminal peptide. It was linked to peptide 'c' which was linked to peptide 'a'. Consequently, the sequence of the protein can be assessed as shown in Figure 2. The complete correct alignment of the sequence of the individual peptides in the mixture data is shown in Figure 3.

It can be pointed out that the identity of the peptides present in the mixture is now defined since all real cleavages occurred on the protein have been established by the above analysis (Figure 2). As can be deduced looking at Figure 3, the sequences of the five unfractionated peptides in the CNBr mixture were the following: EEAM, PSAWPCDECCTRM, IPPRCTCM, DVSPSGCHPACKNCVQTTLGGRDVFWCM and LRIENFCKRRCTPAR. Furthermore, it should be noted that, excepting N-terminal peptide 'b', the starting points of the sequence of remaining individual peptides in the mixture data (5 for peptide 'a' and 10 for peptide 'c' in Figure 3) are consecutive to steps in which E was been identified (4 and 9 in Figure 1B or Figure 3). This trivial observation (E is both the hydrolysis site recognized by protease V8 and C-terminal residue of deriving peptides) allows the user to utilize a quick analysis routine, the

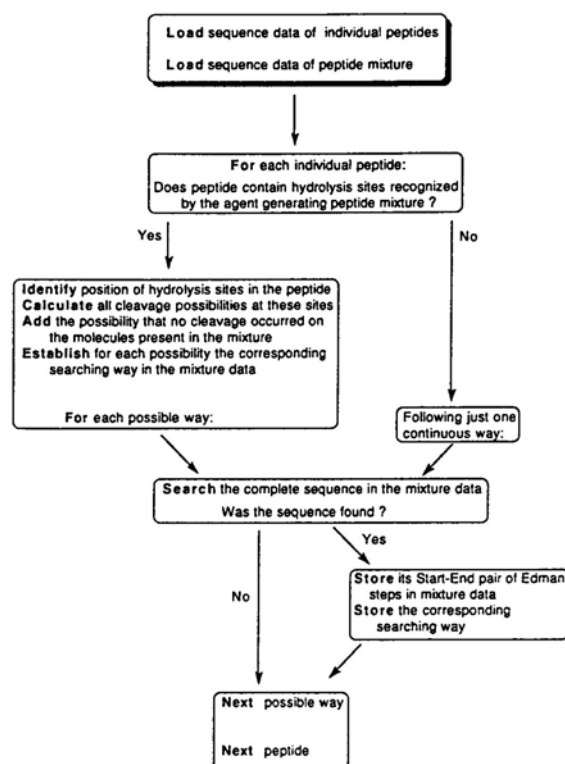


Fig. 4. Basic-like flow chart of the searching algorithm.

search starting just from steps (1, 3, 5, 10 in Figure 1B or Figure 3) simply identified by inputted information about the digesting agent used in obtaining individual peptides. Of course, the quick analysis could fail in identifying some start-end pair if all the expected amino acids present at hydrolysis sites were not correctly inputted. In this case, the full analysis starting from each Edman step can be used. The Basic-like flow chart of the searching algorithm is shown in Figure 4. Sequence data of individual peptides and mixture are kept into memory by matrices of integer variables.

Discussion

The algorithm described in this paper can fail only if the data are not correct owing to some mistake in inputting them or to some wrong interpretation of the sequence analyses which could mainly occur in the case of a very complex peptide mixture. However, using a specific method of hydrolysis producing a limited number of fragments, the correct interpretation of the sequence data just requires a little care even in the case of peptides containing high repetitive sequences. In fact, the carryover of the Edman reaction is a phenomenon that is well

quantified using a modern sequenator. Furthermore, labile residues (e.g. Ser and Thr) partially destroyed during the degradation are simultaneously identified both as PTH-amino acids and PTH- Δ -amino acids. In our experience, the correct interpretation of the sequence data of a mixture containing up to ten peptides (1–2 nmol of each) is not a problem up to 40–50 cycles of degradation.

Not consecutive start–end pairs of Edman steps can be found only in the case of some missing overlap between individual and unfractionated peptides. This occurred when inhibitor WTI was hydrolysed by CNBr and endoproteinase Asp-N. In fact, the protein presents an M-D bond at positions 28–29 (Figure 2) which is equally cleaved by these digesting agents. An uncompleted alignment was obtained using the sequence data of individual CNBr peptides and Asp-N mixture. However, the sequence of the protein could be reconstructed even in this case, since just one overlap was missing. Obviously, the complete sequence could not be assessed if more than one overlap was lacking. In such cases, experimental mixture data obtained by a further hydrolysis method are necessary and the whole sequence of the already aligned peptides can be used as individual sequence data in order to certainly obtain the complete sequence alignment.

Finally, it can be pointed out that the analysis time is dependent on the number of the hydrolysis sites recognized by the agent used to obtain the mixture and contained in the individual sequences rather than the number of single peptides or mixture data. In fact, the sequence of V8 peptide 'b' containing just one M site is searched in CNBr mixture data following two possible ways, while the sequence of V8 peptide 'c' containing three M sites generates eight possibilities when searched in CNBr mixture data (Figure 2). The time required to obtain the results shown in Figure 2 is ~ 25 s using the quick analysis routine.

A method to reconstruct protein sequence by combining sequence data from individual peptides and FAB-MS data from various fragment mixtures has been reported (Petrilli *et al.*, 1991). The approach presented here shows that sequence analysis of peptide mixtures performed on modern automatic sequenators represents an alternative to FAB-MS analysis and presents some advantages. In fact, FAB-MS could fail in identifying all fragments due to the well-documented phenomenon of signal suppression (Naylor *et al.*, 1986); moreover, very small peptides could not furnish any signal. Automatic sequence data are complete and quantitative; for this reason, the development of new algorithms working on such kind of data should be very useful. The present algorithm as well as that already reported (Caporale *et al.*, 1993) should represent a convenient application in utilizing automatic sequence analysis of peptide mixtures. The isolation of

peptides from a single mixture, necessary in utilizing the approach described in this paper, represents the 'charge to be paid' in order to avoid doubts at some sequence position. The program is available from the authors without any contribution. Requests should be accompanied with a 3.5-inch diskette.

Acknowledgements

Research supported by National Research Council of Italy, Special Project RAISA, Sub-project N.2. Paper N. 1903.

References

- Biemann, K. (1980) Amino acid sequence in oligopeptides and proteins. In Waller G.R. (ed), *Biochemical Applications of Mass Spectrometry*. John Wiley, New York, pp. 469–525.
- Cannon, L.E. and Lovins, R.E. (1972) Quantitative protein sequencing using mass spectrometry: computer-aided assembly of protein sequences from N-terminal peptide sequences. *Anal. Biochem.*, **46**, 33–44.
- Caporale, C., Caruso, C., Petrilli, P., Sepe, C., Poerio, E. and Buonocore, V. (1993) A computer program to determine the amino acid sequence of proteins by utilizing data obtained from peptide mixtures. *Protein Seq. Data Anal.*, **5**, 337–344.
- Erickson, B.J. and Jardine, I. (1986) Implementation of the gas chromatographic/mass spectrometric peptide sequencing program PEPALG on a personal computer for off-line analysis. *Biomed. Environ. Mass. Spectrom.*, **13**, 343–346.
- Fairwell, T., Barnes, W.T., Richards, F.F. and Lovins, R.E. (1970) Sequence analysis of complex protein mixtures by isotope dilution and mass spectrometry. *Biochemistry*, **9**, 2260–2267.
- Gray, W.R. (1968) Protein Structure: A new strategy for sequence analysis. *Nature*, **220**, 1300–1304.
- Herlihy, W.C. and Biemann, K. (1981) Advances in gas chromatographic protein sequencing III. Automated interpretation of the mass spectra of the polyamino alcohol derivatives. *Biomed. Mass. Spectrom.*, **8**, 70–77.
- Kitagishy, T., Hong, Y. and Shimonishi, Y. (1981) Computer-aided sequencing of a protein from the masses of its constituent peptide fragment. *Int. J. Peptide Protein Res.*, **17**, 436–443.
- Kitagishy, T., Hong, Y., Takao, T., Aimoto, S. and Shimonishi, Y. (1982) Computation of amino acid sequences of polypeptides from masses of their constituent peptide fragments and amino acid residues released in Edman degradation. *Bull. Chem. Soc. Jpn.*, **55**, 575–580.
- Matsuo, T., Matsuda, H. and Katakuse, I. (1981) Computer program PAAS for the estimation of possible amino acid sequence of peptides. *Biomed. Mass. Spectrom.*, **8**, 137–143.
- Naylor, S., Findeis, A.F., Gibson, B.W. and Williams, D.H. (1986) An approach toward the complete FAB analysis of enzymic digests of peptides and proteins. *J. Am. Chem. Soc.*, **108**, 6359–6364.
- Petrilli, P. (1985) An algorithm for reconstructing protein sequences. *Int. J. Peptide Protein Res.*, **25**, 85–88.
- Petrilli, P. and Colosimo, A. (1990) Computers as tools in protein sequencing. In: Fini, C. and Wittmann-Liebold, B. (eds), *Laboratory Methodology in Biochemistry*. CRC Press, Boca Raton, FL, USA, pp. 129–150.
- Petrilli, P., Sepe, C. and Pucci, P. (1991) A new procedure for peptide alignment in protein sequence determination using fast atom bombardment mass spectral data. *Biol. Mass. Spectrom.*, **20**, 115–120.
- Poerio, E., Carrano, L., Garzillo, A.M. and Buonocore, V. (1989) A trypsin inhibitor from the water-soluble protein fraction of wheat kernel. *Phytochemistry*, **28**, 1307–1311.
- Poerio, E., Caporale, C., Carrano, L., Caruso, C., Vacca, F. and Buonocore, V. (1994) The amino acid sequence and reactive site of a single-headed trypsin inhibitor from wheat endosperm. *J. Prot. Chem.*, **13**, 187–194.
- Shimonishi, Y., Hong, Y., Katakuse, I. and Hara, S. (1981) A new method for protein sequence analysis using Edman-degradation, field-desorption mass spectrometry and computer calculation. Sequence determination of the N-terminal CNBr fragment of Streptomyces erythraeus lysozyme. *Bull. Chem. Soc. Jpn.*, **54**, 3069–3075.

Received on February 26, 1994; accepted on July 21, 1994