A computer program to determine the amino acid sequence of proteins by utilizing data obtained from peptide mixtures

Carlo Caporale¹, Carla Caruso¹, Pasquale Petrilli², Ciro Sepe¹, Elia Poerio¹ and Vincenzo Buonocore¹.

¹ Dipartimento di Agrobiologia ed Agrochimica, Università' della Tuscia, via S.Camillo de Lellis, 01100, Viterbo, Italy, FAX N. 0761-357242.

² Istituto di Industrie Agrarie, Università' di Napoli, Portici, Italy.

Abbreviations: PyEt-, pyridylethyl-; FAB-MS, Fast atom bombardment Mass spectrometry; PTH, phenylthiohydantoine; RP-HPLC, reverse phase high-performance liquid chromatography

Enzymes: Endoproteinase Lys-C; Endoproteinase Asp-N

Correspondence to: C. Caporale
Abstract

We formulated a computer program with the aim of determining the primary structure of proteins by utilizing sequence data obtained from peptide mixtures. The protein under study is digested with different methods and the unfractionated peptide mixtures are submitted to automatic sequence analysis. The data are cross-related to make exact peptide overlaps according to the strategy proposed by Gray since 1968. In order to clarify possible doubts at some positions of the sequence, due to the simultaneous overlap of more than one residue, the output of the program also furnishes the masses (MH+\(^{+}\)) of all possible peptides utilized for the sequence determination for each mixture. Thus, if existing, the doubts can be clarified performing a fast atom bombardment mass spectrometry (FAB-MS) analysis on the unfractionated peptide mixture(s). To test the validity of the proposed method, the complete amino acid sequence of a trypsin inhibitor from wheat kernel (coded WTI) was redetermined; no experimental FAB-MS data were necessary in the presented case.
Introduction

Since 1968 Gray proposed a strategy for determining protein sequences limiting peptide purification steps. The strategy consists in utilizing phenylthiohydantoine (PTH) amino acid data of the sequence analyses performed on the unfractonated peptide mixtures obtained by digesting the protein with different methods [1]. Obviously, several PTH-amino acid residues are identified at each step of the Edman degradation by using this approach. The example Gray proposed to explain this strategy is shown in Fig. 1, where the amino acids released by ten steps of Edman degradation performed on CNBr (cleaving after Met) and clostripain (cleaving after Arg) digests of bovine pancreatic ribonuclease are listed [1]. It should be evident that, starting from the third step in the CNBr set and from the first step in the clostripain set, the unambiguous partial sequence Arg-Asn-Leu-Thr-Lys is compatible with both sets of PTH data. Similarly, starting from the sixth and ninth step in CNBr set and from the first step in clostripain set the partial sequences Arg-Glu-Thr-Gly-Ser and Arg-Cys are consistent with both sets of data, respectively. Moreover, several partial sequences showing doubts at some positions due to the overlap of more than one residue can be deduced. By means of 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, Gray concluded that the complete sequence of ribonuclease (124 residues) could be deduced if it were possible to specifically cleave the protein after Arg, Lys, Cys, Met, Tyr, His and to identify the first ten residues of each peptide [1].

Gray's strategy inspired some methods which use data obtained from peptide mixtures to determine the amino acid sequence of a protein [2-10]. All these methods, revised in [11], require various kind of data combined each other such as the quantitative determination of the amino acids in the mixtures by isotopic dilution [2, 3], the mass and the partial N-terminal sequences of the fragments generated
by the proteolytic agents [4], the amino acid composition and the N- and C-terminal sequences of the protein [5] and the qualitative identification of the PTH-amino acids [6, 7]. Furthermore, a method based on the gas chromatographic mass spectrometric analysis of mixtures containing small peptides generated by means of mild acid and/or nonspecific endopeptidase hydrolysis has been reported [8-10].

With the advent of the modern pulsed-liquid phase sequencers equipped on-line with PTH analyzers, the original Gray's idea can be applied to real cases. In this paper we report an improved version of the Gray's algorithm implemented on a personal computer. The only data required are the PTH amino acid residues identified by sequence analyses performed on the unfractonated peptide mixtures deriving from the protein digestions. The complete sequence of a trypsin inhibitor from wheat (71 residues) [12, 13] was reconstructed by analyzing simultaneously the data deriving from three different methods of hydrolysis.

Materials and methods

*Triticum aestivum*, pure variety San Pastore, was kindly supplied from Istituto Nazionale per la Cerealicoltura (Rome, Italy). Pulsed liquid-phase automatic sequencer (model 477A) equipped on-line with PTH analyser (model 120A) and relative reagents were from Applied Biosystems. HPLC procedures were carried out on a Beckmann GOLDF™ apparatus equipped with a variable-wavelength monitor model 166. Sequencing grade endoproteinase Lys-C and endoproteinase Asp-N were from Boehringer Mannheim Italia SpA. The C8 column (Supelcosil LC-308 0.46 x 25 cm) was from Supelchem (Milano, Italy). All other reagents were of analytical grade.

Purification of inhibitor WTI
Inhibitor WTI was purified as previously described [12]. The homogeneity of the preparation was evaluated by reverse-phase high-performance liquid chromatography (RP-HPLC) on a C8 column.

Determination of protein concentration

The protein concentration was determined by the Bio-Rad assay, following the manufacturer's instructions and using bovine serum albumin as a standard.

Reduction and pyridylethylation

Inhibitor WTI (75 nmol) was treated with dithiothreitol (7.5 μmol) in 0.2 ml of 0.25 M Tris-HCl buffer pH 8.0, 2 mM EDTA, 6 M guanidine-HCl, for 2 h at 37 °C under nitrogen atmosphere. Alkylation was carried out with 4-vinylpyridine (400 μmol) for 2 h at room temperature under nitrogen atmosphere in the dark. The reaction was stopped by freezing and reagents in excess were removed by RP-HPLC on the C8 column. Aliquots of the pyridylethylated protein (PyEt-WTI) were freeze-dried separately.

Digestion with endoproteinase Lys-C

PyEt-WTI (10 nmol) was digested with endoproteinase Lys-C (enzyme/substrate 1:50 w/w) in 100 μl of 25 mM Tris-HCl buffer, pH 8.5, 1 mM EDTA for 8 h at 37 °C. The reaction was stopped by freezing. The mixture was desalted by filtration on a Sep-pak C18 cartridge eluted with trifluoroacetic acid 0.1% in acetonitrile/water 70/30 v/v and then lyophilized.

Digestion with endoproteinase Asp-N

PyEt-WTI (10 nmol) was digested with endoproteinase Asp-N (enzyme/substrate 1:500 w/w) in 100 μl of 50 mM sodium phosphate, pH 8.0 for 8 h at 37 °C. The reaction was stopped by freezing. The mixture was desalted as described in the previous section and lyophilized.
Cleavage with CNBr

PyEt-WTI (10 nmol) was cleaved by cyanogen bromide [14]. The reaction was carried out in 0.2 ml of 70% formic acid for 24 h at room temperature under nitrogen atmosphere in the dark. The mixture was then diluted with water (2 ml) and lyophilized twice. The freeze-dried sample was dissolved in water (0.2 ml), desalted as previously described and lyophilized.

Sequencing peptide mixtures

Peptide mixtures obtained by CNBr, endoproteinase Lys-C and endoproteinase Asp-N treatments were dissolved in aqueous 0.1% trifluoroacetic acid (50-100 µl). A portion of each fragment mixture (1-2 nmol of the original protein) was submitted to automatic sequence analysis. The analyses were interrupted when no PTH residue was identified for at least 3 cycles of Edman degradation.

Hardware and software

Programs were written in Applesoft Basic and implemented on an Apple II GS equipped with an Image Writer printer and a 3.5 inches disk drive. The operative system was Apple Prodos. The software is compatible with all personal computers Apple (series II) equipped with a minimum of 128 K RAM. A version for Apple Macintosh computers is currently in progress.

Results and discussion

Sequencing data

The validity of the strategy proposed for determining protein primary structures from sequence data obtained from peptide mixtures has been tested on a trypsin
inhibitor, coded WTI, whose sequence has been determined recently by a classical approach [13]. PyEt-WTI was separately digested by using three different methods chosen on the basis of WTI amino acid composition [12]. In table 1 are reported pmoles data of all the PTH-amino acids obtained, for each step of the Edman degradation, sequencing the mixture deriving from the digestion with endoproteinase Lys-C. The results obtained from the Asp-N digestion are reported in table 2, whereas table 3 shows the results deriving from the chemical cleavage with CNBr. As it can be observed, several PTH-amino acids were easily identified at each of 40, 22 and 28 steps of the Edman degradation, respectively. It should be noted that homoserin and homoserin lacton arise from cleavage with CNBr of protein containing Met. The retention time of PTH-homoserin in the HPLC system of analysis of PTH-amino acids is very similar to that of PTH-Thr and, consequently, PTH-Met is not identifiable in the data of the CNBr set (table 3).

The most critical remark done to the Gray's strategy comes from the supposed difficult interpretation of HPLC chromatograms of the PTH-amino acids deriving from peptide mixtures due to the carryover of the Edman reaction and to the identification of labile residues (e.g. Ser and Thr) partially destroyed during degradation [11]. In our experience these problems have been solved with the advent of modern pulsed-liquid phase sequencers equipped on-line with PTH analyzers. In fact, the identification of a number of residues at each step of degradation needs just a little of care. The carryover is a phenomenon that can be quantified and labile PTH residues are easily identified (e.g. Ser and Thr are simultaneously identified both as PTH-amino acids and PTH-Δ-amino acids). Furthermore, due to the possible simultaneous presence of the same residue(s) at various sequencing steps (e.g. Cys was the only residue identified at the 7th step in the CNBr set, table 3) and to the degradation of the short peptides which is completed in few steps, the interpretation of the chromatograms is quite simplified (e.g. just one sequence was identified from the 17th to 28th step in the CNBr mixture, table 3, and from the 25th to 40th step in the Lys-C mixture, table 1). Of
course, the cleavage methods to be chosen should preferentially produce large fragments that are more suitable for automatic sequencing.

The residues identified at each sequencing step from the three mixtures (tables 1-3) are reported in Fig. 2. They represent the data necessary to determine the sequence of the protein.

Algorithm

The program is divided into the following three sections:

A) Keyboard input/edit section: all the data are inputted from the keyboard and stored on the disk. The user will input all the amino acids identified at each step of Edman degradation performed on all the digested samples. Furthermore, some additional information about the digesting agents must be stored; in particular the user will indicate all the expected amino acids present at sites of hydrolysis on the basis of the specificity of the hydrolytic agent, their position (N- or C- terminal) in the produced peptides and, finally, if their PTH derivatives are identifiable in the sequence analysis of the corresponding mixture (e.g. PTH-Met is not identifiable in the mixture deriving from the CNBr treatment). Moreover, the number of total amino acid residues determined by the amino acid analysis of the protein (74 in the present case [12]) is inputted as an additional information to evaluate if the sequence to be determined reached a satisfactory length.

B) Searching section: the elucidation of the searching algorithm we used to assess the sequence of the protein WTI can be summarized as shown in Fig. 3. Once loaded in memory the data of the three mixtures, the search of the residue(s) common to all sets in consecutive steps of degradation started from the first cycle in the three sets. Consequently, the starting position of the pointers indicating the correct alignment of the sets of data was 1. Obviously, the N-terminal sequence of the protein was determined in this way. In fact, the only residue common to all the first steps was E (Fig. 2). Extending the search to the next steps, E at the second, A at the third, M at the fourth (M was common to the sets Lys-C and Asp-N and
not identifiable in the set CNBr) and P at the fifth steps were the only consecutive residues common to all sets. Since no further common residue was found at the sixth steps, the unambiguous sequence EEAMP of five residues could not be further extended. It aborted and was stored in a result file. This extension was obtained not considering the presence of hydrolysis sites ("free" extension). However, M being a site of hydrolysis of the protein cleaved with CNBr, the sequence EEAMP was truncated at site M and the possibility of extension of the sequence EEAM of four residues was analyzed. Now, the starting positions of the pointers for the correct alignment of the sets of data were the following: 5 (1+4: previous starting position + extension length) for the set Lys-C, 5 (1+4) for the set Asp-N and 1 for the set CNBr as EEAM was the complete sequence of the N-terminal peptide deriving from the CNBr treatment and the search in the CNBr set started from the first residue of another peptide. Moreover, an additional important information was stored. In fact, the protein was hydrolyzed at position 4 by CNBr treatment and the mass (MH+ ) of the peptide EEAM present in the CNBr mixture and corresponding to the sequence 1-4 was calculated. It should be noted that the algorithm is not dependent from the effectiveness of the digesting agent. In fact, if the hydrolysis at position 4 did not occur, the further extension of the sequence EEAM should have been correctly found by the above "free" extension. The algorithm always verifies if the cleavage occurred when an extension including an hydrolysis site is found; of course, the same extension will be found twice if a partial cleavage at an hydrolysis site occurred and the corresponding PTH data were inputted in both ways. The unambiguous free extension PSAWPCCD was found starting from the Edman steps indicated by the above pointers. It should be noted that the residue D was common to the sets Lys-C and CNBr only, as D is an hydrolysis site of endoproteinase Asp-N which leaves D residues at the N-terminus of the produced peptides. Consequently, the sequence PSAWPCCD of 8 residues was truncated before D and the possibility of extension of the sequence EEAMPSAWPCC of 11 residues (4+7) was analyzed after the
calculation of the new starting pointers. The sequence position 11 of the new cleavage was stored and the mass of the peptide 1-11 present in the Asp-N mixture was calculated. The sequence reached the length of 71 residues by progressive extensions, until no further extension was found. The sequence aborted and was stored in the result file with all the data regarding the cleavages deriving from the three digestions. As it can be observed, no ambiguity was found in 70 of 71 positions. The only doubt regarded the position 66, as both C and R were common to the three sets after the correct alignment of data determined by the starting pointers. The flow chart of the searching algorithm is shown in Fig. 4.

C) Output section: the results obtained by the searching algorithm are shown in Fig. 5. The output sequence of 71 residues was identical to that determined by a classical approach in which R was found at position 66 [13]. The output also indicates the starting pointers used when no further extension is found. This can be of some utility in verifying the inputted data if the sequence does not reach a satisfactory length. As previously anticipated, the masses of all possible peptides matching the sequence are calculated for each mixture. It should be evident that the doubt at position 66 could be clarified performing a FAB-MS analysis just on one of the three mixtures. However, no experimental FAB-MS data was necessary in this case. The amino acid present at position 66 is R as the absence of free cysteine residues in inhibitor WTI [12] is not compatible with the presence of a thirteenth Cys residue at this position. In fact, all the Cys residues (12) involved in disulfide bridges are unambiguously located in the sequence.

A similar result in determining the sequence of inhibitor WTI should be obtained by using the FORTRAN program PROSEQ2 [6,7], which needs molecular weight data of the peptides present in the utilized mixtures combined with data of the amino acids released from successive steps of Edman degradation. The algorithm we have designed and implemented on a personal computer is quite different since mass data are not required and the use of mass spectrometry represents an optional tool to speed-up the determination of the sequence unambiguously. In
fact, the doubts at some positions of the sequence can be clarified by supplying additional data obtained by a further method of hydrolysis and/or by the sequence analysis of the whole protein if the uncertainties regard the N-terminal region. Moreover, the C-terminal peptide can be easily isolated by modern techniques and sequenced, if the uncertainties are contained in this region.

It should be noted that this approach is obviously more suitable in assessing the sequence of small proteins rather than large ones. In fact, some difficulties in the identification of PTH amino acid residues could arise in the sequence analysis of very complex mixtures deriving from the digestions of a large protein. In such case, the original complex mixtures can be analyzed by the present method after a separation step yielding more simple mixtures; of course, many hydrolysis methods should be used to clarify the uncertainties. A total of 3000 PTH data can be kept in memory by a three-dimensional matrix of integer variables and simultaneously analyzed by the algorithm we designed. The time required to obtain the results shown in Fig. 5 from the analysis of the inputted data is about 1.5 min on an Apple II GS.

In conclusion, we propose a flexible method based on the original Gray's idea which furnishes many information about the sequence of a protein very quickly, even if the determination of the complete and unambiguous sequence in more complex cases needs additional data that can be chosen by the user on the basis of the output of the program. The program is available from the authors without any contribution. Requests should be accompanied with a free 3.5 inches diskette.

Acknowledgements

Research supported by National Research Council of Italy, Special Project RAISA, Sub-project N.2. Paper N. XXX and by a grant of "Progetto Finalizzato:
Resistenze genetiche delle piante agrarie agli stress abiotici e biotici" from Ministero Agricoltura e Foreste.

References

Legend to tables and figures

Table 1. Sequence analysis performed on the PyEt-WTI digest obtained by endoproteinase Lys-C treatment. Pmoles data of PTH amino acids identified at each of 40 Edman degradation steps are underlined. Cys residues were identified as PTH-PyEt-Cys.

Table 2. Sequence analysis performed on the PyEt-WTI digest obtained by endoproteinase Asp-N treatment. Pmoles data of PTH amino acids identified at each of 22 Edman degradation steps are underlined. Cys residues were identified as PTH-PyEt-Cys.

Table 3. Sequence analysis performed on the PyEt-WTI peptide mixture obtained by chemical cleavage with CNBr. Pmoles data of PTH amino acids identified at each of 28 Edman degradation steps are underlined. Cys residues were identified as PTH-PyEt-Cys.

Fig. 1. Gray's strategy. Amino acids that would be released by ten steps of Edman degradation performed on A) CNBr and B) clostripain digests of bovine pancreatic ribonuclease. The unambiguous partial sequences Arg-Asn-Leu-Thr-Lys, Arg-Glu-Thr-Gly-Ser and Arg-Cys are consistent with both sets of data.

Fig. 2. Amino acids identified at each sequencing step performed on PyEt-WTI peptide mixtures obtained by A) endoproteinase Lys-C B) endoproteinase Asp-N and C) CNBr treatment.

Fig. 3. Progressive determination of WTI sequence by the searching algorithm. The positions in the sequence of the hydrolysis sites were calculated for each peptide mixture.
Fig. 4. Flow chart of the searching algorithm.

Fig. 5. Output of the program. The identified sequence of inhibitor WTI is indicated.

The masses (MH⁺) of all possible deduced peptides present in the mixtures were calculated. The molecular weight of Cys residues was calculated as molecular weight of PyEt-Cys residues. The molecular weight of the C-terminal residue of Met in peptides of the CNBr mixture was calculated as molecular weight of both a) homoserin and b) homoserin lacton.