Journal of Peptide Science
J. Peptide Sci. **11**: 225–234 (2005)
Published online 1 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/psc.595

JOURNAL OF PEPTIDE SCIENCE

On peptide *de novo* sequencing: a new approach‡

RENATO BRUNI,[a,b]* GIANLUIGI GIANFRANCESCHI[a,c] and GIORGIO KOCH[a,b]

[a] PolyDART: Data Analysis Research Team for Polymers, 03015 Fiuggi (FR), Italy
[b] Department of Computer and Systems Science, University of Roma 'La Sapienza', 00185 Roma, Italy
[c] Department of Cellular and Molecular Biology, University of Perugia, 06123 Perugia, Italy

Received 5 May 2004; Accepted 10 June 2004

Abstract: A procedure is presented for the automatic determination of the amino acid sequence of peptides by processing data obtained from mass spectrometry analysis. This is a basic and relevant problem in the field of proteomics. Furthermore, it has an even higher conceptual and applicative interest in peptide research, as well as in other connected fields. The analysis does not rely on known protein databases, but on the computation of all amino acid sequences compatible with the given spectral data. By formulating a mathematical model for such combinatorial problems, the structural limitations of known methods are overcome, and efficient solution algorithms can be developed. The results are very encouraging both from the accuracy and computational points of view. Copyright © 2004 European Peptide Society and John Wiley & Sons, Ltd.

Keywords: combinatorial optimization; *de novo* sequencing; mass spectrometry; peptide analysis

INTRODUCTION

The analysis of the amino acid sequence of peptides and proteins is one of the most important and common issues in biological and medical research. This is especially true after the conclusion of the *Genome Project*, which contributed to question the idea of a unique correspondence between genes and the generated protein, opening the doors to the *Proteoma Project*. Indeed, 'if the 1990s were the decade of genomics, the first 10 years of the new century are set to become the decade of proteomics' [1]. A basic analysis in proteomics is the identification of proteins, that is their amino acid sequence, and of any possible changes of that sequence which might have occurred due to alternative splicing, post-translational modifications (glycosylation, phosphorylation, acetylation, methylation, etc.), or other possible causes. Protein analyses are generally achieved by dividing a protein molecule into its component peptides (via enzymatic digestion and subsequent fractionation with HPLC or capillary electrophoresis), and by analysing such peptides. Thus, peptide sequencing arises as a fundamental step in protein identification. Following this, homology and alignment methods exist to build up the full sequence of a protein from the sequences of its peptides [2,3].

Besides being part of the protein identification problem, peptide sequencing has an importance of its own in a number of situations such as the study of unknown peptides, research for new drugs and the synthesis of peptide-like active factors and peptides used in therapy (a number of hormones are peptides). Therefore, sequencing clearly represents an essential issue in the field of peptide research, as well as in other connected fields.

Peptide sequencing was initially achieved by biochemists by means of the Edman method [4,5], which may be implemented either manually or through the use of automatic instrumentation (protein sequenators). However, such a procedure has several drawbacks: in particular, it is difficult to apply whenever the material to be sequenced is only available in small quantities, or in the case where the *N*-terminus of the peptide is blocked, or even when a mixture of peptides with similar chemico-physical features are being studied, so that they cannot be separated easily [6]. Furthermore, the Edman method calls for a subjective experienced check of the resulting pattern against available databases.

On the other hand, mass spectrometry is now a widely used and well-established approach to peptide sequencing. When applied to peptides or protein molecules, such a technique gives the exact molecular weight of the full molecule, as well as those of its fragments possibly produced during the analysis [7–10]. The study of the weight pattern in the spectrum provides clues toward understanding the peptide sequence [11–14]. The sequencing can be helped further by the use of the so-called mass spectrometry/mass spectrometry (MS/MS, or tandem mass) methodology [15–19]. According to this procedure, a precursor ion, usually the protonated peptide generated by means of a suitable ionization method, is selected and collided with non-reactive gas molecules. This interaction leads to the fragmentation of the selected ion, and the collision-generated decomposition products undergo mass analysis. Therefore, all the fragments under analysis refer

* Correspondence to: Dr Renato Bruni, Dipartimento di Informatica e Sistemistica, Università di Roma 'La Sapienza', Via Michelangelo Buonarroti 12-00185 Roma, Italy; e-mail: bruni@dis.uniroma1.it

‡ Italian Patent number: MI2002A 000396. International Patent Application number: PCT/IB03/00714

to just the selected precursor ion. These experiments can be performed using different instrumental configurations, mainly triple quadrupole (QQQ), quadrupole time-of-flight (Q-TOF) and ion trap devices. Note that by using the second approach accurate mass measurements of either precursor or fragment ions can be obtained (so that their elemental composition can be determined), while the third method allows multiple, sequential collision experiments (MS^n) to be performed.

A typical MS/MS spectrum, however, does not contain any direct reference to amino acids, being a mere succession of peaks corresponding to different molecular weights. Further analysis is then required, and generally performed as follows. To begin with, all peaks below a certain intensity are removed, being too noise-affected to be considered significant. After this, the higher molecular weight is assumed to be that of the full peptidic complex under analysis, the others corresponding to fragments of the peptide complex. The problem of extracting such significant peaks (hereinafter called the peptide pattern) out of the whole of the spectral data is usually dealt with by some heuristic procedure (not described here) based on some *a priori* knowledge of the fragmentation process. As a matter of fact, peptides normally fragment in a predictable manner and luckily a major portion of these fragments forms discrete ion series directly related to the peptide sequence.

A classical peptide analysis uses MS/MS spectrometry patterns and checks for peptide-specific weight patterns (peptide tags, or fragment fingerprints) against similar patterns available from databases or virtually generated from known sequences by some fragmentation model or algorithm [15,17,20]. In this case candidate patterns are ranked by their distance from, or probability to fit with, the experimental ones. The whole procedure relies on data reported in the available databases [21–23].

The use of databases clearly assumes that the protein (or the peptide) under investigation belongs to an already known set. But if this is not the case, or if our protein differs from a standard known form by the sequence undergoing one of the above mentioned modifications, alternative methods are required and direct identification must be addressed. Direct peptide sequencing (also known as '*de novo* sequencing') is achieved by various recently available techniques. These procedures either look for continuous sequences of *N*-terminal and/or *C*-terminal fragments differing by just one amino acid, which is therefore identified, or iteratively generate a large number of virtual sequences and evaluate the match of the corresponding (theoretical) mass patterns with the (actual) mass pattern of the peptide under investigation.

In both cases, the sequence can be obtained uniquely when the pattern contains the complete series of fragments. This, however, is often unlikely to occur. Indeed, the peptide usually does not break at every conjunction of amino acids, and if the intensity of the hitting is increased, the peptide may break at locations which are not a conjunction of amino acids. This makes the problem very difficult for existing *de novo* techniques.[§] We note that a non-unique, or a partial, peptide sequence may still be considered satisfactory if the final goal is solely the identification of the protein within a known database, possibly turning to an expert system or to some action by the human operator, to overcome the lack of information.

Therefore, the peptide sequencing problem can arise in a case of protein identification or as a problem on its own, and it has quite different features in the two cases.

In the case of protein identification:

- there is always a database in the background;
- the task often is to acknowledge whether a given protein coincides with that usually synthesized by a specific tissue, or significantly differs from it because of some pathological behaviour, and in what percentage over a wide population;
- there is, as a consequence, the need to process a large number of mass spectra, to scan spectral data over various ranges, and find high speed analysis methods;
- possibly non-unique peptide sequences are ranked (statistically) according to their fit to the data.

On the other hand, if peptide sequencing is the central issue, then:

- there is no well-established database;

[§] Information about available softwares can be found at the following web sites:
–Fragment fingerprints:
BayMatch; Micromass: http://www.micromass.co.uk
Mascot; Matrix Science: http://www.matrixscience.com
Mass Search; ETH: http://cbrg.ethz.ch
Mowse; Human Genetics Res. Center: http://www.seqnet.dl.ac.uk
MS-Tag, MS-Fit, Ms-Seq; Univ. California, San Francisco: http://prospector.ucsf.edu
PepFrag, ProFound; Rockefeller University: http://prowl.rockefeller.edu
Pep Sea; Protana: http://www.protana.com
PeptideSearch; EMBO: http://www.mann.embl-heidelberg.de
SEQUEST; Univ. Washington, Seattle: http://fields.scripps.edu/sequest
Turbo SEQUEST; Thermofinnigan: http://www.thermo.com
Sonar MS/MS; Proteometrics: http://www.proteometrics.com
–*De novo* sequencing:
DeNovoX; Thermofinnigan: http://www.thermo.com
Mass Seq; Micromass: http://www.micromass.co.uk
PEAKS, Bioinformatics Solutions Inc.: http://www.bioinformaticssolutions.com
Spectrum Mill; Agilent: http://www.agilent.com
–Fragment alignment:
BLAST; EMBO: http://www.mann.embl-heidelberg.de.
PatternHunter; Bioinformatics Solutions Inc.: http://www.bioinformaticssolutions.com

- it is essential to identify the whole peptide sequence (or all possible sequences) with the highest possible accuracy;
- algorithms are therefore required to be accurate, rather than fast;
- scanning is used to reduce background noise;
- ranking of possible solutions is in terms of internal features (such as entropy, or likeliness of configurations).

Therefore a strong need exists for an analysis procedure for peptide sequencing that overcomes the drawbacks and limits of the above mentioned methods, that is aimed at both solving the protein identification problem and plainly and accurately determining peptide sequences as such. Note that there could always be cases when the information contained in the spectrum is simply not sufficient to determine a unique sequence, because more than one sequence exists which perfectly fits such a spectrum.

MATERIALS AND METHODS

An innovative analysis procedure for the determination of all possible sequences of a peptide by analysing data obtained from raw mass spectrometry analysis is described. This was obtained by developing a mathematical model of the problem and by searching for all possible sequences of given components satisfying certain constraints. Such constraints were formalized and mathematically expressed. Due to the strong combinatorial nature of the problems, the search for the solutions was carried out by means of specialized branching techniques.

Development of the Mathematical Model

The mathematical model was defined by a set of decision variables, having values within a given domain, and by a set of constraints on those variables. By denoting the number of possible amino acids by n (e.g. 20), the set of indices corresponding to such amino acids in increasing weight order by $A = \{1, 2, \ldots, n\}$, the (unknown) number of amino acid molecules contained in the analysed peptide by m, and the set of indices corresponding to such amino acid molecules ordered from the N-terminus to the C-terminus by $B = \{1, 2, \ldots, m\}$, the following set of binary (or Boolean) variables was used, with $i \in A$ and $j \in B$

$$x_{ij} = 1 \quad \text{if the } i\text{-th amino acid is in position } j\text{-th of the peptide, 0 otherwise.}$$

These decision variables must be related through a set of constraints. The structure of such constraints contains the *a priori* knowledge of the fragmentation process (Figure 1), while the numerical values are given by the available mass spectrometry data.

Each peak of weight w in the spectrum can be due to the presence of one of the various 'classical' types of fragment (a-ion, b-ion, c-ion, x-ion, y-ion, z-ion) having a weight w, or to the presence of 'additional' fragmentation (losses of small neutral molecules such as water, ammonia, carbon dioxide, carbon monoxide, or breaking of a lateral chain; see also [24]) of one of the above types of fragments having a weight greater than w, or to the presence of a multicharged fragment having a weight of a multiple of w, or even to the presence of some spurious component having a weight w. One should also consider the presence of some noise peaks, even if they generally have low intensity values. Altogether, therefore, the scenario is tricky.

As is customary, all peaks below a certain intensity must be removed, together with all peaks that cannot correspond to any possible 'classical' fragment. After this step, the resulting sequence of peaks is considered, as follows. The heaviest peak, i.e. the observed weight of the overall peptide complex, is denoted by MH^+ and by p_k all other remaining peaks, $k \in P$, which are assumed to be the weights of the fragments. The molecular weight of the i-th amino acid is denoted by Maa_i. Only for the purpose of speeding up the constraint checking operation, molecular weights both for amino acids and peptides will be approximated by integer numbers. As a matter of fact, our results in peptide sequencing show that such precision is largely adequate. There are, of course, no theoretical impediments to the use of a higher numeric precision in the described procedure. The constraint securing compatibility with the overall weight of the complex has the following structure (see e.g. [24,25]).

$$1 + 18 + \sum_{i \in A, j \in B} [x_{ij}(Maa_i - 18)] = MH^+$$

A generic subsequence of B is denoted by S, and a constant is denoted by c_k whose value is -27 for a-ions, 1 for b-ions, 18 for c-ions, 45 for x-ions, 19 for y-ions, 3 for z-ions, 36 for y-NH_3-ions, 37 for y-H_2O-ions, etc. The constraints securing compatibility with the weight of the various types of fragments introduced above are therefore:

$$\exists S \subseteq B \text{ such that}$$
$$c_k + \sum_{i \in A, j \in S} [x_{ij}(Maa_i - 18)] = p_k \quad \forall k \in P$$

Finally, constraints imposing that the sequence has exactly one amino acid for each position j of the peptide are of the type:

$$\sum_{i \in A} x_{ij} = 1 \quad \forall j \in B$$

Since b-ions and y-ions are by far the most common kind of fragment, only those types of fragments are searched for in the spectrum. Our results confirm such an assumption. The search can obviously be extended to other types of fragments, but this would require more computational time. Further information, such as the presence or absence of some amino acid, may sometimes be available. This corresponds to additional constraints on the values of the variables that are easily taken into account so as to speed up the search.

It is of interest to identify all solutions of the described constrained problem. Due to the possible presence of the above mentioned 'unusual' fragment, it may frequently happen that a set of constraints does not admit any feasible solution. For this

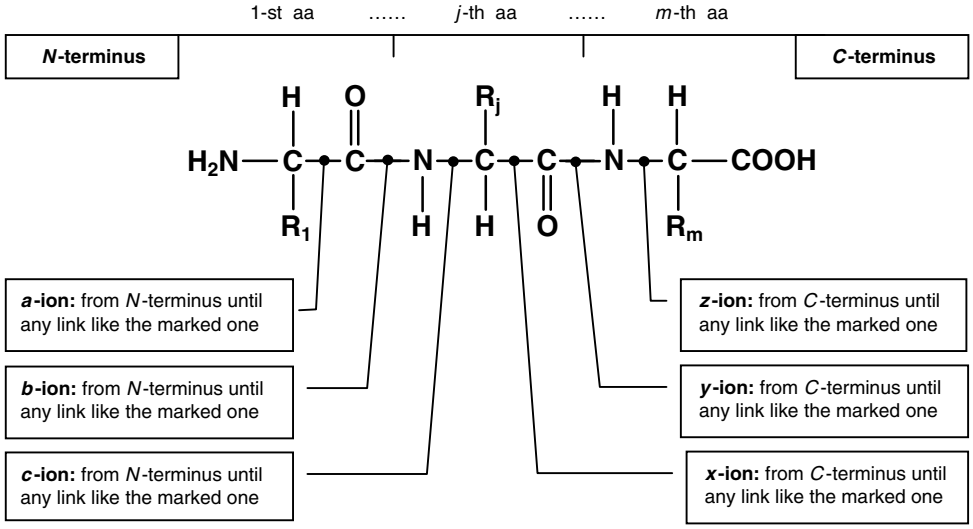


Figure 1 Classical fragmentation of a peptide chain. Protonation at the amine followed by cleavage at the peptide amide bond leaves an ammonium ion, called the ***y*-ion**, which is a '*C*-terminal fragment' because the charge is retained on the *C*-terminus. Other *C*-terminal fragments are the ***x*-ions** and ***z*-ions**, which are, however, less common. When instead the charge is retained on the carbonyl, an acylium ion is formed, called the ***b*-ion**, which is an '*N*-terminal fragment'. Other *N*-terminal fragments are ***a*-ions** and ***c*-ions**, which are again less common. Each ion has a weight depending on its components, according to the following criteria: a ***b*-ion** weighs 1 plus the sum of its component amino acids, each of which is decreased by 18; an ***a*-ion** weighs the same as the corresponding ***b*-ion** minus 28; a ***c*-ion** weighs the same as the corresponding ***b*-ion** plus 17; a ***y*-ion** weighs 19 plus the sum of its component amino acids, each of which is decreased by 18; an ***x*-ion** weighs the same as the corresponding *y*-ion plus 26; a ***z*-ion**, finally, weighs the same as the corresponding *y*-ion minus 16. Moreover, when a fragment retains more than one charge, the observed weight is a fraction of the actual ion weight.

reason, the procedure accepts a value t called the *mismatch number*. An amino acid sequence is defined by a *solution* to the sequencing problem if, and only if, for each peak p_k, with $k \in P$, except at most t values, there is one of the above defined constraints which is verified.

The Solution Algorithm

In order to obtain such solutions, search techniques based on *branching* are used [26]. Such techniques rely on systematic and recursive partitioning of the space in regions which are easier to explore. This is achieved by progressively *fixing* values for the x variables within their binary domain $\{0,1\}$, thus generating subproblems with progressively decreasing dimensions. The search evolution may then be represented through a *search tree*. Each node of the branch tree corresponds to a partial solution, given by the fixings performed on the path from the root of the search tree up to the current node.

However, for the described problems, the number of possible vectors x is enormous, and exponentially increases with the size of the problem [27], as would the time needed to generate them all (for example, there are more than 100 billion different amino acid sequences having the same weight of 1000 D). To speed up the listing, those search tree branches that do not yield solutions are not explored. This means checking whether the current branch of the search tree corresponds to a partial solution not respecting more than t constraints. Since such a check can be computationally heavy when considering the whole set of constraints, only some of them are generated and tested, under suitable conditions, at each node of the search tree, thus developing an innovative specialized algorithm inspired by techniques of *delayed row generation* [28].

This algorithm has been conveniently implemented in the C++ language and runs on a standard PC. Solution times are of the order of tenths of seconds for complexes up to $MH^+ = 1000$ D. Heavier complexes may be solved in a short time in those cases when some additional information is available (for instance, the weights of all possible fragments, or the composition of some fragments). The solutions are ordered lexicographically by the molecular weights of the components, modulo a permutation, thus allowing an easy search in the case where the number of possible solutions is large.

RESULTS

The results of analysis on raw MS/MS spectra are presented in Figures 2 and 3. In the former case, the information contained in the spectrum is sufficient to determine a unique sequence. In the latter case, on the contrary, the spectrum does not contain enough information for the determination of a unique sequence, and by progressively increasing the number of acceptable mismatches four solutions are found. Figure 4 shows how, by increasing the quantity of available information (i.e. by considering more peaks of the spectrum), the number of possible sequences decreases until a unique sequence is obtained. Further experimental results for sequencing problems are briefly reported and commented in the Experimental section below.

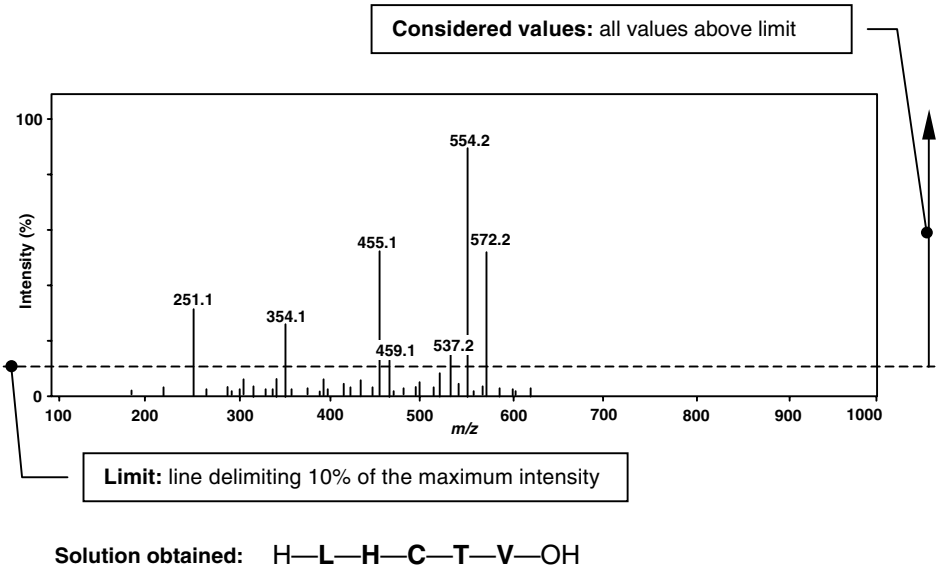


Figure 2 Analysis of a typical MS/MS spectrum, where all peaks above 10% of the maximum intensity are considered. The precise limit value that should be used depends on the experiment, and an exact rule cannot be given. The list of considered values is: 251.1, 354.1, 455.1, 459.1, 537.2, 554.2, 572.2. The analysis of such data gives, at zero mismatch, the unique sequence H-Leu-His-Cys-Thr-Val-OH.

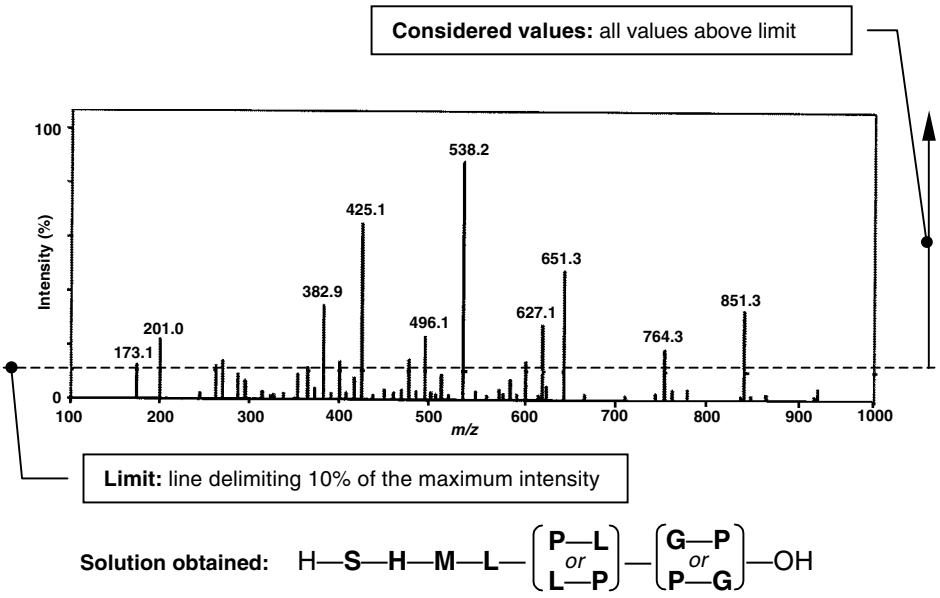


Figure 3 Analysis of a typical MS/MS spectrum, where all peaks above 10% of the maximum intensity are considered. The list of considered values is therefore: 173.1, 201.0, 262.1, 270.0, 286.1, 366.2, 382.9, 401.8, 425.1, 479.2, 496.1, 515.0, 496.1, 538.2, 611.1, 627.1, 651.3, 764.3, 851.3. The analysis of such data does not gives a solution until nine mismatches. When ten mismatches are used, the obtained sequence is H-Ser-His-Met-Leu-(Pro-Leu *or* Leu-Pro)-(Gly-Pro *or* Pro-Gly)-OH, that altogether means four possible unique sequences.

DISCUSSION

The proposed analysis system therefore presents a number of novel points.

(a) The procedure does not look for an amino acid sequence that is closest to the given fragment mass spectrum (according to some *a priori* chosen distance criterion), but rather for sequences which fit it exactly.

(b) The procedure is able to deal with situations where the spectrum does not contain enough information for an unequivocal determination of the sequence. In this case, all possible sequences that fit the spectrum are listed, with equal *'dignity'* and in a lexicographic order.

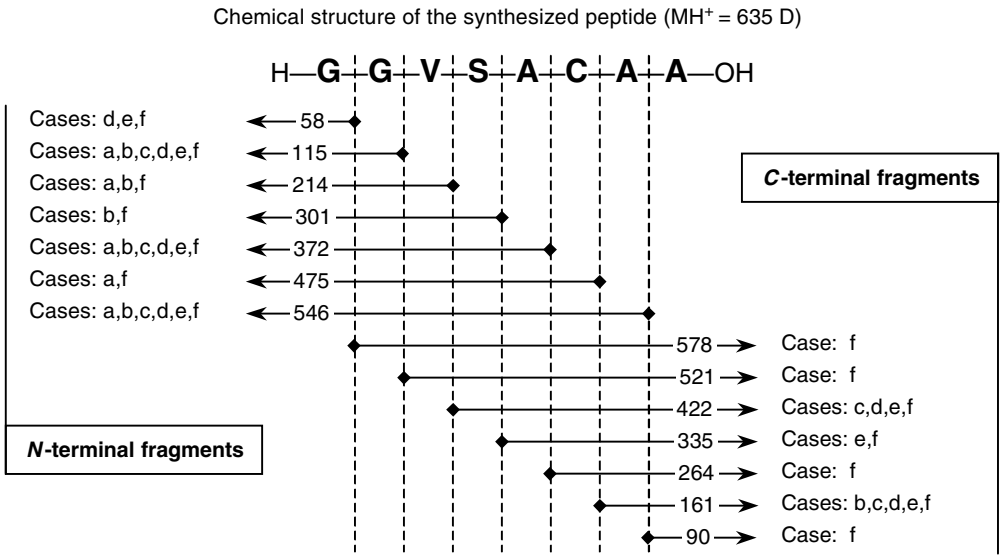


Figure 4 By progressively considering more fragments, the number of possible solutions decreases. The following cases were considered: **a:** 115, 214, 372, 475, 546, which leads to eight possible solutions; **b:** 115, 161, 214, 372, 546, leads to eight possible solutions; **c:** 115, 161, 372, 422, 546, leads to eight possible solutions; **d:** 58, 115, 161, 372, 422, 546, leads to four possible solutions; **e:** 58, 115, 161, 335, 372, 422, 546, leads to one solution; **f:** 58, 90, 115, 161, 214, 264, 301, 335, 372, 422, 475, 521, 546, leads to one solution.

(c) The search for a peptide sequence does not use databases, nor expert systems; rather, it follows a formal mathematical procedure based on a branching algorithm and internally processes the raw experimental spectrum.
(d) The procedure is based on computationally efficient branching techniques, in order to tackle the heavy computational time requirement. This is not a minor issue, since even for molecules with a weight of about 500 D there are several thousand corresponding amino acid combinations.
(e) The detailed peptide sequencing achieved by the procedure allows the identification of possible modifications in the peptide itself. It suffices to include the weight of any additional component to the basic amino acid weights.
(f) The algorithm accounts for spurious components of the spectrum due, for instance, to 'unusual' breaking down of the amino acid chain, multicharged fragments, satellite peaks or peaks too close to each other.
(g) The algorithm easily incorporates (and in so doing, it advantageously saves time since it reduces the number of possible sequences) any other information that happens to be available, such as for instance: the known presence and/or absence of given amino acids, or of given subsequences of them; the known *N*- or *C*- origin of a given fragment; results from previous sequencing attempts.
(h) For a relatively large peptide, the number of possible sequences identified by our method may well turn out to be very large if the fragmentation from MS/MS analysis has been incomplete. However, this number may be considerably decreased (until just one sequence results) by reprocessing the largest fragments by means of the proposed procedure.

A few additional comments are worthy of further research work:

(a) The number of significant digits used to denote the weight location of the mass spectrometry peaks deserves some attention. It is known that the equivalent mass involved in the bindings leads to non-integer values for the amino acid weights. In addition, the existence of different isotopes introduces averaging problems. Indeed, there are weight tables that report up to six significant decimal digits. The basic question, however, is whether the accuracy and the stability of the actual analysis equipment is able to tackle such detailed numerical representation in a significant way.
(b) In the case where no solution is found by using a low mismatch number, one possible strategy is to add other components to the basic elements to be identified in the peptide sequence. An alternative strategy might be to leave a number of mismatched peaks out of the first analysis, and subsequently to resubmit them to further analysis in order to possibly identify them, for instance, as terminals of different types, or as peaks belonging to the pattern of a different molecule.
(c) The procedure used to extract the considered peptide pattern from the spectrum has to be carefully designed. One might cut everything which

falls below a certain threshold, or a certain percentage of the highest peak, or even adopt different strategies over different weight ranges. But one should keep in mind that cutting an informative peak as noisy leads to a loss of information and possibly to non-uniqueness of the achieved sequence. On the contrary, assuming as informative a peak brought into the considered pattern by mere noise may lead to a radical (erroneous) modification of the sequence arrived at.

EXPERIMENTAL

Lyophilized samples were solubilized in methanol and injected for electron-spray mass spectrometry analysis. Mass spectrometer: LCQ-MS THERMOQUEST/ESI-ION trap, except where stated otherwise. Capillary temperature 220 °C, capillary voltage 10 V, spray voltage 4 KV, collision energy as stated. Some examples of the application of the proposed procedure follow.

Table 1

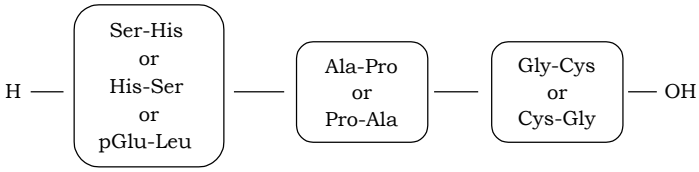
General Description

Peptide fraction isolated from wheat sprout chromatin, active in the control of HL60 cell proliferation.
From the MS/MS spectrum of native and unknown peptide sequence with MH^+ 571, two prominent fragments with m/z 393 and 347 have been obtained. Obviously, as two fragments cannot represent the fragmentation at every peptide bond, in this case it is impossible to obtain a unique sequence.
Collision energy 19 Kev

Considered Mass Peaks

571 393 347

Results



Comments

Since the number of considered fragments is exceedingly poor, the number of possible solutions is large (12). However, they all are related, in the sense that they are all obtainable by various combinations of the above three boxes.

Table 2

General Description

Peptide isolated from bovine seminal plasma and involved in the control of steroidogenesis *in vitro*.
Collision energy 19 KeV

Considered Mass Peaks

859 748 599 484 397 282 261 133

Results

Comments

The result shows four possible sequences:
(a) pGlu-Val-Ala-Asp-Ser-Asp-Gln-Asn-OH
(b) pGlu-Ala-Val-Asp-Ser-Asp-Gln-Asn-OH
(c) pGlu-Val-Ala-Asp-Ser-Asp-Lys-Asn-OH
(d) pGlu-Ala-Val-Asp-Ser-Asp-Lys-Asn-OH
The sequences (c) and (d) may be discarded because the amino acid analysis of the native fraction following acid hydrolysis demonstrated the absence of Lys. The sequences (a) and (b) have been synthesized. Preliminary experiments show that the biological activity *in vitro* of the peptide with sequence (a) is quite similar to that exerted by the native peptide fraction.

Table 3

General Description
H-Gly-Gly-Leu-Phe-Gly-Gly-Ala-Gly-OH synthetic peptide
Collision energy 17 KeV

Considered Mass Peaks
635 521 432 375 261 228 147

Results
H-Gly-Gly-Leu-Phe-Gly-Gly-Gly-Ala-OH
H-Gly-Gly-Leu-Phe-Gly-Gly-Ala-Gly-OH
H-Gly-Gly-Leu-Phe-Gly-Gly-Gln-OH
H-Gly-Gly-Leu-Phe-Gly-Gly-Lys-OH
H-Asn-Leu-Phe-Gly-Gly-Gly-Ala-OH
H-Asn-Leu-Phe-Gly-Gly-Ala-Gly-OH
H-Asn-Leu-Phe-Gly-Gly-Gln-OH
H-Asn-Leu-Phe-Gly-Gly-Lys-OH

Comments
The analysis gives eight solutions because the fragmentation is not complete and the molecular weight of some amino acids (when inside the peptide chain, i.e. decreased by 18) may be the sum of two other amino acids:
Asn(114) = Gly(57) + (57);
Gln(128) or Lys(128) = Gly(57) + Ala(71).

Table 4

General Description
H-Gly-Gly-Leu-Phe-Gly-Gly-Ala-Gly-OH synthetic peptide
Collision energy 23 KeV

Considered Mass Peaks
635 578 560 521 432 404 375 345 261 228 147

Result
H-Gly-Gly-Leu-Phe-Gly-Gly-Ala-Gly-OH

Comments
The higher collision energy produces more fragments. The analysis of such data with zero mismatches does not give a solution. With two mismatches a unique solution is obtained.

Table 5

General Description
Peptide fraction purified from bovine spermatozoa.
Collision energy 19 KeV

Considered Mass Peaks
913.3 766.1 748.0 720.3 637.1 633.1 524.2 505.0 477.2 435.1 409.3 407.0 390.2 281.0 277.1 166.0

Results
H-Phe-Glu-Leu-Asp-Gly-Ala-Asp-Phe-OH
H-Phe-Glu-Leu-Asp-Ala-Gly-Asp-Phe-OH
H-Phe-Glu-Leu-Ser-Gly-Val-Asp-Phe-OH
H-Phe-Glu-Leu-Ser-Val-Gly-Asp-Phe-OH
H-Phe-Glu-Leu-Ser-Arg-Asp-Phe-OH
H-Phe-Glu-Leu-Asp-Gln-Asp-Phe-OH
H-Phe-Glu-Leu-Asp-Lys-Asp-Phe-OH

Comments
The analysis of these data did not give any solution until two mismatches. With three mismatches seven solutions are obtained. It is noteworthy that this set of solutions shows homologous amino acid sequences at both *N*-terminus (H-Phe-Glu-Leu-) and *C*-terminus (-Asp-Phe-OH).

Table 6

General Description

A Tyr phosphorylated peptide (from [29]). This peptide contains two modified amino acids: acetylated aspartic acid (Ac-Asp) and phosphotyrosine (Tyr(P)): Ac-Asp-Tyr(P)-Val-Pro-Met-Leu-OH.
The molecular weights of Ac-Asp (158) and Tyr(P) (243) have therefore been added to the possible peptide components.
Mass spectrometer: BRUKER ESQUIRE-LC ESI QIT (see [29]).

Considered Mass Peaks

859.1 728.3 702.2 597.3 500.2 401.1 360.3

Results

Ac-Asp-Gly-Gly-Glu-Val-Pro-Met-Leu-OH
Ac-Asp-Gly-Glu-Gly-Val-Pro-Met-Leu-OH
Ac-Asp-Glu-Gly-Gly-Val-Pro-Met-Leu-OH
Ac-Asp-Gly-Ala-Asp-Val-Pro-Met-Leu-OH
Ac-Asp-Gly-Asp-Ala-Val-Pro-Met-Leu-OH
Ac-Asp-Ala-Gly-Asp-Val-Pro-Met-Leu-OH
Ac-Asp-Ala-Asp-Gly-Val-Pro-Met-Leu-OH
Ac-Asp-Asp-Gly-Ala-Val-Pro-Met-Leu-OH
Ac-Asp-Asp-Ala-Gly-Val-Pro-Met-Leu-OH
Ac-Asp-Gly-Ser-Val-Val-Pro-Met-Leu-OH
Ac-Asp-Gly-Val-Ser-Val-Pro-Met-Leu-OH
Ac-Asp-Ser-Gly-Val-Val-Pro-Met-Leu-OH
Ac-Asp-Ser-Val-Gly-Val-Pro-Met-Leu-OH
Ac-Asp-Val-Gly-Ser-Val-Pro-Met-Leu-OH
Ac-Asp-Val-Ser-Gly-Val-Pro-Met-Leu-OH
Ac-Asp-Ala-Ala-Thr-Val-Pro-Met-Leu-OH
Ac-Asp-Ala-Thr-Ala-Val-Pro-Met-Leu-OH
Ac-Asp-Thr-Ala-Ala-Val-Pro-Met-Leu-OH
Ac-Asp-Ser-Arg-Val-Pro-Met-Leu-OH
Ac-Asp-Arg-Ser-Val-Pro-Met-Leu-OH
Ac-Asp-Asn-Glu-Val-Pro-Met-Leu-OH
Ac-Asp-Glu-Asn-Val-Pro-Met-Leu-OH
Ac-Asp-Asp-Gln-Val-Pro-Met-Leu-OH
Ac-Asp-Gln-Asp-Val-Pro-Met-Leu-OH
Ac-Asp-Asp-Lys-Val-Pro-Met-Leu-OH
Ac-Asp-Lys-Asp-Val-Pro-Met-Leu-OH
Ac-Asp-Tyr(P)-Val-Pro-Met-Leu-OH

Comments

Although the number of solutions appears to be large, they all show only two kinds of sequences:
(a) Ac-Asp-X-X-Val-Pro-Met-Leu-OH: 26 solutions
(b) Ac-Asp-Tyr(P)-Val-Pro-Met-Leu-OH: one solution.
The sequences (a) have been obtained because the molecular weight of Tyr(P) may be also the sum of the molecular weights of two or three amino acids. If the presence of one residue of Tyr(P) is postulated, only sequence (b) is obtained.

Acknowledgements

We are grateful to Professor L. De Angelis and Mr F. Giavarini (Laboratory of Mass Spectrometry, Department of Pharmaceutical Sciences, University of Milan) for the experimental mass spectrometry analyses. We are also in debt to the unknown reviewers for their contribution to the improvements of the final version of the paper.

REFERENCES

1. Anderson NL, Matheson AD, Steiner S. Proteomics: applications in basic and applied biology. *Curr. Opin. Biotech.* 2000; **11**: 408–412.
2. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 1993; **233**: 123–138.
3. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci.* 1998; **7**: 445–456.
4. Edman P. A method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* 1950; **4**: 283–293.
5. Heinrikson RL. The Edman degradation in protein sequence analysis. In *Biochemical and Biophysical Studies of Proteins and Nucleic Acids*, Lo TB (ed). Elsevier: New York 1984; 285–302.
6. Jensen ON, Podtelejnikov AV, Mann M. Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal. Chem.* 1997; **69**: 4741–4750.
7. Lee TD. Fast atom bombardment and secondary ion mass spectrometry of peptides and proteins. In *Methods of Protein Microcharacterization*, Shively JE (ed). Humana Press: Clifton, NJ, 1986; 403–441.
8. Siuzdak G. *Mass Spectrometry for Biotechnology*. Academic Press: New York 1996.
9. Stults JT. Peptide sequencing by mass spectrometry. *Method Biochem. Anal.* 1990; **34**: 145–201.
10. Bradley CV, Williams DH, Hanley MR. Peptide sequencing using the combination of Edman degradation, carboxypeptidase digestion and fast atom bombardment mass spectrometry. *Biochem. Biophys. Res. Commun.* 1982; **104**: 1223–1230.
11. Chou DK, Evans JE, Jungalwala JB. Identity of nuclear high-mobility-group protein, HMG-1, and sulfoglucuronyl carbohydrate-binding protein, SBP-1, in brain. *J. Neurochem.* 2001; **77**: 120–131.
12. Taka H, Kaga N, Fujimura T, Mineki R, Imaizumi M, Suzuki Y, Suzuki R, Tanokura M, Shindo N, Murayama K. Rapid determination of parvalbumin amino acid sequence from *Rana catesbeiana* (pI 4.78) by combination of ESI mass spectrometry, protein sequencing, and amino acid analysis. *J. Biochem.* 2000; **127**: 723–729.
13. Lin T, Glish GL. *C*-Terminal peptide sequencing via multistage mass spectrometry. *Anal. Chem.* 1998; **70**: 5162–5165.
14. Damer CK, Partritge J, Pearson WR, Haystead TA. Rapid identification of protein phosphatase 1-binding proteins by mixed peptide sequencing and data base searching. Characterization of a novel holoenzymic form of protein phosphatase 1. *J. Biol. Chem.* 1998; **273**: 24 396–24 405.
15. Taylor JA, Johnson RS. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 1997; **11**: 1067–1075.

16. Taylor JA, Johnson RS. Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal. Chem.* 2001; **73**: 2594–2604.
17. Johnson RS, Taylor JA. Searching sequence databases via *de novo* peptide sequencing by tandem mass spectrometry. *Methods Mol. Biol.* 2000; **146**: 41–61.
18. Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. *De novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 1999; **6**: 327–342.
19. Schlosser A, Lehmann WD. Patchwork peptide sequencing: extraction of sequence information from accurate mass data of peptide tandem mass spectra recorded at high resolution. *Proteomics* 2002; **2**: 524–533.
20. Li KW, Jimenez CR, Van Veelen PA, Geraerts WP. Processing and targeting of a molluscan egg-laying peptide prohormone as revealed by mass spectrometric peptide fingerprinting and peptide sequencing. *Endocrinology* 1994; **134**: 1812–1819.
21. Mackey AJ, Haystead TAJ, Pearson WR. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell. Proteomics* 2002; **1**: 139–147. Software FASTA, FASTS, FASTF, University of Virginia, Charlottesville http://www.virginia.edu.
22. Tabb DL, McDonald WH, Yates JR III. DTA Select and contrast: tools for assembling and comparing protein identification from shotgun proteomics. *J. Proteome Res.* 2002; **1**: 21–26. Software DTA Select, Contrast, University of Washington, Seattle http://fields.scripps.edu/sequest.
23. Ngoka LCM, Gross ML. Location of alkali metal binding sites in endothelin A selective receptor antagonists, cyclo(D-Trp-D-Asp-Pro-D-Val-Leu) and cyclo(D-Trp-D-Asp-Pro-D-Ile-Leu), from multistep collisionally activated decompositions. *J. Mass Spectrom.* 2000; **35**: 265–276.
24. Mancinelli L, Chillemi F, Cardellini E, Marsili V, Giavarini F, De Angelis L, Lugaro G, Gianfranceschi GL. Molecular models of acidic peptides from pea bud chromatin and seminal plasma. Divalent cations-mediated interaction with DNA. *Biol. Chem.* 1999; **380**: 31–40.
25. Mancinelli L, Lugaro G, De Angelis L, Gianfranceschi GL. Mass spectral and electrophoretic characterization of acidic peptides bound to chromatin of pea bud. *Mol. Biol. Rep.* 1998; **25**: 163–172.
26. Hooker JN. *Logic Based Methods for Optimization*. Wiley: New York 2000.
27. Garey MR, Johnson DS. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman: New York 1979.
28. Nemhauser GL, Wolsey LA. *Integer and Combinatorial Optimization*. Wiley: New York 1988.
29. Zolodz MD, Wood KV. Detection of tyrosine phosphorylated peptides via skimmer collision-induced dissociation/ion trap mass spectrometry. *J. Mass Spectrom.* 2003; **38**: 257–264.

