An Innovative Approach to Peptide De Novo Sequencing*

Renato Bruni [1,2], Gianluigi Gianfranceschi [1,3], Giorgio Koch [1,2]

1 PolyDART Data Analysis Research Team for Polymers, Via Prenestina 96, 03015 Fiuggi (FR), Italy
2 Dept. of Computer and System Science, University of Roma "La Sapienza", 00185 Roma, Italy
3 Dept. of Cellular and Molecular Biology, University of Perugia, 06123 Perugia, Italy

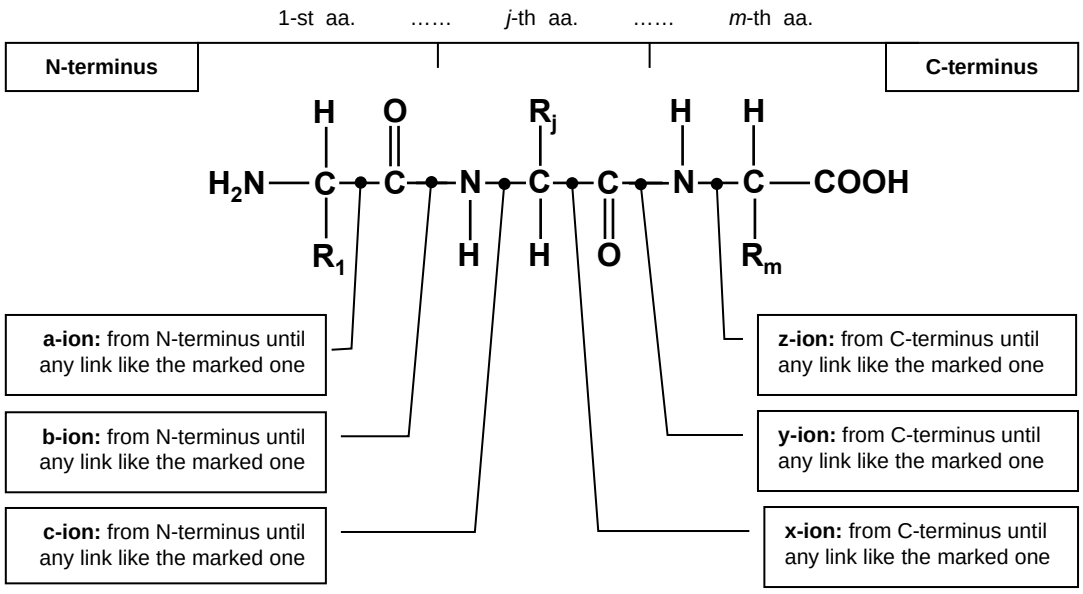
* Italian Patent number: MI2002A 000396. International Patent Application number: PCT/IB03/00714.

Introduction

The analysis of the amminoacidic sequence of a peptide is a basic and fundamental issue in Proteomics. Such analysis is generally performed by using Mass Spectrometry equipment. This procedure exhibits several advantages on other methods. In particular, a mass/mass (or tandem mass) analysis is very often used. This latter technique produces a spectrum displaying the molecular weight of the full molecule and those of its fragments produced during the analysis itself.
A typical mass/mass spectrum, however, does not contain any direct reference to amminoacids, being a mere succession of molecular weight peaks. Further analysis of such spectral data is then requested.

The Fragmentation Process

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 weight *w*, or due 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, etc.) of one of the above types of fragments having weight greater than *w*, or due to the presence of a multicharged fragment having weight of a multiple of *w*, or even due to the presence of some spurious component having weight *w*. One should also consider the presence of possible noise peaks, even if they generally have low intensity values. Altogether, therefore, the scenario is tricky.



The Proposed Approach

Our approach to the sequencing problem does not rely on known protein databases, nor does it look for protein-specific weight patterns of the obtained spectrum (mass fingerprints) or peptide-specific weight patterns (peptide tags, or fragment fingerprints), nor does it rank candidate patterns by their distance from, or probability to fit with, the experimental ones. Rather, it directly carries on computation of all amminoacidic sequences compatible with the given (mass) spectrum data. This is a combinatorial problem, whose solution calls for mathematical models and efficient algorithms.

The Mathematical Model

We first remove, as customary, all peaks below a certain intensity, and, moreover, all peaks which cannot correspond to any possible "classical" fragment. After this the resulting sequence of peaks is considered, as follows. Denote, as usual, by MH^+ the heaviest peak, i.e. the observed weight of the overall peptidic complex, and by p_k all other remaining peaks, $k \in P$, which are assumed to be the weights of the fragments. Denote, moreover, by Maa_i the molecular weight of the i-th amminoacid, by n the number of possible amminoacids (e.g. 20), by $A = \{1,2,\ldots,n\}$ the set of indices corresponding to such amminoacids in increasing weight order, by m the (unknown) number of amminoacidic molecules contained in the analyzed peptide, and by $B = \{1,2,\ldots,m\}$ the set of indices corresponding to such amminoacidic molecules ordered from the N-terminal to the C-terminal, the following set of binary (or Boolean) variables is used, with $i \in A$ and $j \in B$.

x_{ij} = 1 if the i-th amminoacid is in position j-th of the peptide, 0 otherwise

Said decision variables must be related through a set of constraints. The structure of such constraints contains our a priori knowledge of the fragmentation process, while the numerical values are given by available mass spectrometry data. The constraint securing compatibility with the overall weight of the complex has the following structure.

$$19.023 + \sum_{i \in A, j \in B} [\, x_{ij} \,(\, \mathrm{Maa}_i - 18.015)] = \mathrm{MH}^+$$

Denote now by S_k a generic subsequence of B, and by c_k a constant whose value is –27.002 for *a*-ions, 1.008 for *b*-ions, 18.039 for *c*-ions, 45.017 for *x*-ions, 19.023 for *y*-ions, 3.000 for *z*-ions, etc. Clearly, all the above numerical values should be rounded according to the numerical precision of considered spectrometry data. Note, however, that in every case the theoretical procedure remains the same. The constraints securing compatibility with the weight of the various types of fragments introduced above are therefore:

$$\exists\ S_k \subset B \quad \text{such that} \quad c_k + \sum_{i \in A, j \in S} [\, x_{ij} \,(\, \mathrm{Maa}_i - 18.015)] = p_k \quad \forall\ k \in P$$

Finally, constraint imposing that the sequence has exactly one amminoacid for each position j of the peptide are of the type:

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

We now have a mathematical formulation of a constrained problem, all the solutions of which are possible sequences that exactly match the analyzed spectrum. 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 reason, the procedure accepts a value t called *mismatch number*. An amminoacid sequence is defined 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.

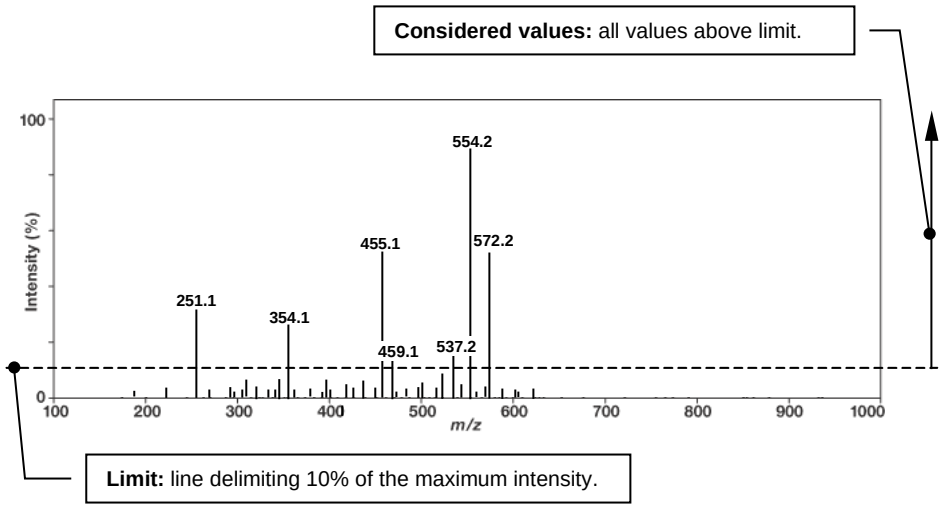
The Solution Algorithm

In order to obtain such solutions, search techniques based on *branching* are used. Such techniques rely on systematic and recursive partitioning of the space in regions which are easier to be explored. This is achieved by progressively *fixing* values for the x variables within their binary domain {0,1}, thus generating subproblems with progressively decreasing dimension. 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 enormously large, and exponentially increasing in the size of the problem. So would therefore be the time needed to generate all of them. (In order to give an idea, there are more then 100 billions of different amminoacids sequences having the same weight of 1000 D). In order to speed up the listing, those search tree branches which do not yield solutions are not to be 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 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*.

This algorithm has been conveniently implemented in 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 anyway be solved in 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). Found solutions are ordered lexicographically by the molecular weights of the components, modulo a permutation, thus allowing an easy search in the case the number of possible solutions is large.

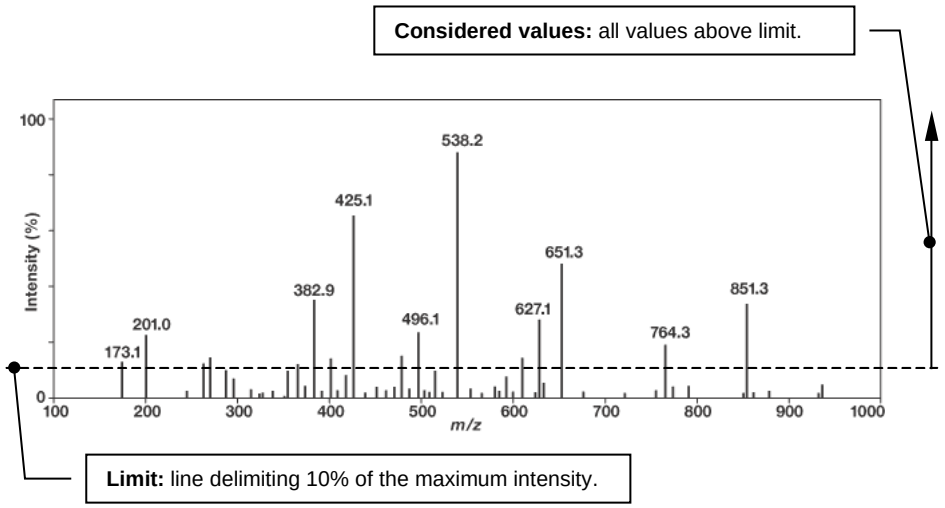
Example 1



Solution obtained: H—**L**—**H**—**C**—**T**—**V**—OH

Analysis of a typical mass/mass spectrum, where all peaks above 10% of the maximum intensity are considered. The precise limit value that should be used depends on the experiment. 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 founds, at 0 mismatch, the unique sequence H-Leu-His-Cys-Thr-Val-OH.

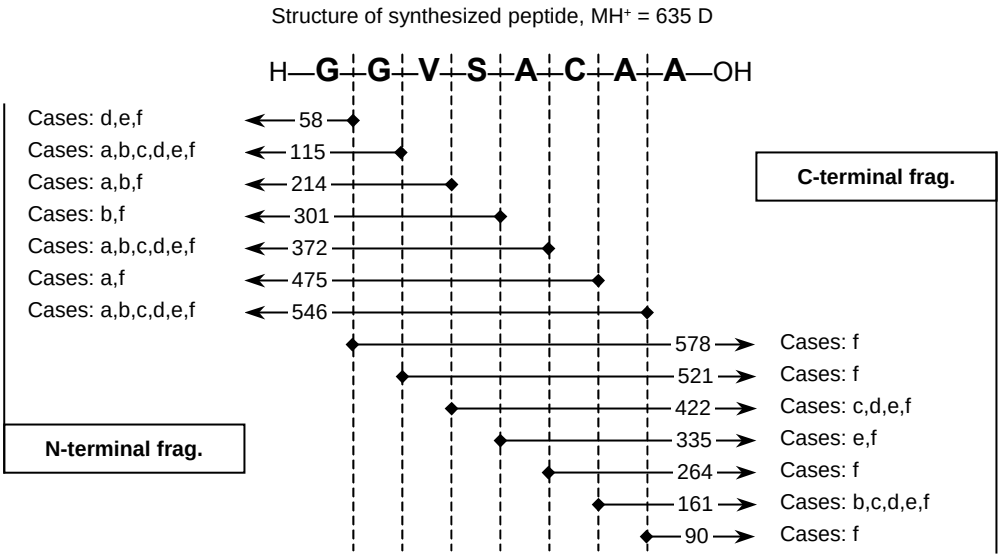
Example 2



Solution obtained: H—**S**—**H**—**M**—**L**—(**P**—**L** *or* **L**—**P**)—(**G**—**P** *or* **P**—**G**)—OH

Analysis of a typical mass/mass 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 solution until 9 mismatches. When using 10 mismatches the obtained sequence is: H-Ser-His-Met-Leu-(Pro-Leu *or* Leu-Pro)-(Gly-Pro *or* Pro-Gly)-OH, that altogether means 4 possible unique sequences.

Example 3



By progressively considering more fragments, the number of possible solution decreases. We considered the following cases: **a:** 115, 214, 372, 475, 546, which leads to 8 possible solutions; **b:** 115, 161, 214, 372, 546, leads to 8 possible solutions; **c:** 115, 161, 372, 422, 546, leads to 8 possible solutions; **d:** 58, 115, 161, 372, 422, 546, leads to 4 possible solutions; **e:** 58, 115, 161, 335, 372, 422, 546, leads to 1 solution; **f:** 58, 90, 115, 161, 214, 264, 301, 335, 372, 422, 475, 521, 546, leads to 1 solution.

Discussion

The analysis system we propose appears to be new under a number of issues.

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

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

c) The search for a peptide sequence does not use data bases, 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 with 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 thousands of corresponding amminoacidic combinations.

e) The detailed peptide sequencing achieved by the procedure allows the identification of possible modifications in the peptide itself.

f) The algorithm accounts for spurious components of the spectrum due for instance to "unusual" breaking down of the amminoacid chain, multicharged fragments, satellite peaks or peaks too close each other.

g) The algorithm easily incorporates (and by doing that it advantageously saves time since it reduces the number of possible sequences) any other information that happens to be available, such as for instance: known presence and/or absence of given amminoacids; 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 mass/mass 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.