A Combinatorial Optimization Approach for Determining Composition and Sequence of Polymers*

Renato Bruni
Dep. of Electronic and Information Engineering,
University of Perugia, Via G. Duranti, 93 - 06125 Perugia - Italy.
E-mail: renato.bruni@diei.unipg.it

Abstract

Polymers are compounds formed by the joining of smaller, often repeating, units linked by covalent bonds. The analysis of their structure is a fundamental issue in a number of fields. This work gives an exact mathematical formalization of the problems of determining the composition or the sequence of polymers by processing data obtained from their tandem mass spectrometry analysis and describes effective solution algorithms for such problems. The procedure is exemplified by considering the case of peptides, but may be used for generic polymeric compounds submitted to mass spectrometry. The analysis does not rely on databases, but on computation of solutions compatible with the given spectral data. Note that the proposed approach guarantees finding all the above solutions, while other known methods cannot give this guarantee. Both the computational running times and the biological accuracy of the experimental analyses are encouraging.

Keywords: Integer Programming, Mass Spectrometry, Peptide or Polymer Analysis

1 Introduction

Polymeric compounds are formed by the joining of smaller units, here generically called *components*, linked by covalent bonds. The analysis of polymeric compounds is one of the most important and frequent issues in biological and medical research, as well as in several other applicative fields. A particularly relevant example of polymer analysis is constituted by the case of *peptide* analysis. Peptides are the short polymeric molecules constituting all the proteins

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

and are usually constituted by a single sequence of components called *amino acids*. The two extremes of the sequence are respectively called N-terminus and C-terminus. This sequence may sometimes be closed in a cycle, obtaining a so-called *cyclic* peptide. The analysis consists in finding the *combination* of components (i.e. which amino acids compose the peptide) or, even better, the *sequence* of components (i.e. which amino acid and in which position of the peptide, since peptides having the same composition but different sequences may have very different behaviors).

A first approach to peptide analysis was the so-called Edman method [7], that may be implemented either manually or through the use of automatic devices called protein sequenators. However, such a procedure exhibits several drawbacks [10]. Therefore, as in the case of many other polymers, the approach which is currently prevailing consists in the use of *mass spectrometry* analysis (e.g. [14, 16]). Such technique can provide the absolute molecular weight distribution of a number of molecules in the form of a *spectrum*. The study of the weight pattern in the spectrum can be used to understand the structure of such molecules, especially when the analysis is further supported by the so called mass spectrometry/mass spectrometry (MS/MS, or tandem mass) methodology (e.g. [21]). When applied to peptides, this procedure works as follows. After the first mass analysis, some molecules of the protonated peptide under analysis, called *precursor ion*, are selected and collided with non reactive gas molecules. This interaction leads to the fragmentation of many of such molecules, and the collision-generated decomposition products undergo the second mass analysis. Therefore, such analysis can provide the absolute molecular weight of the full precursor ion, as well as those of the various ionized fragments obtained from that precursor ion. Note that the presence of these fragments constitutes the only source of information about the inner structure of the peptide under analysis: in absence of fragmentation, the inner structure would be unknown. Though fragmentation is a stochastic process, some types of fragments, called *standard* fragments, are more common then others. Such standard fragments can be of six different types, called a, b, c, x, y, z, as described in detail in Section 4. Fragments are ionized by retaining one or more electrical charges. Non ionized fragments, on the contrary, do not appear in the spectrum. This analysis technique can be applied to other polymeric compounds and can be carried out by using several instrumental configurations, mainly triple quadrupole (QQQ), quadrupole time-of-flight (Q-TOF) and ion trap devices.

This work gives an exact mathematical formalization of the problems of determining the composition or the sequence of a generic polymeric compound. It moreover presents an innovative procedure for the automatic solution of such relevant problems. Such a procedure is exemplified by considering the case of peptides, which constitute a particularly relevant example, but may be used for other generic polymers. The above is obtained by developing innovative mathematical models of the two problems, as explained in Section 4, and by searching for:

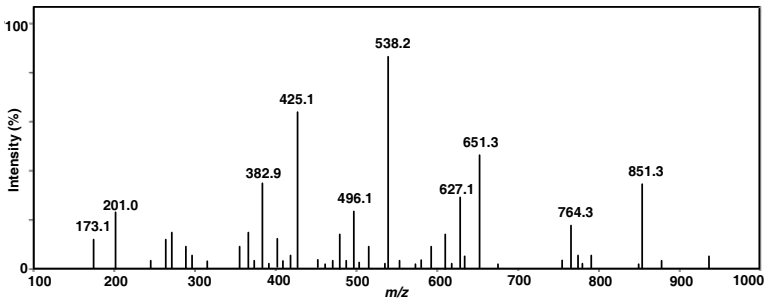


Figure 1: A MS/MS spectrum generated by collision-induced dissociation.

- all possible combinations of components compatible with the given MS/MS spectrum, by finding all solutions satisfying a set of constraints;
- all possible sequences of components compatible with the given MS/MS spectrum, by finding all solutions satisfying another set of constraints.

The proposed models have some similarities with the multidimensional integer knapsack problem [15] and represent an evolution of the one proposed in [3]. Due to the strong combinatorial nature of such problems, the search for the solutions is carried on by means of specialized *branching* techniques, developed as described in Section 5. Structural limitations (concerning type of fragments and type of charges) of other known approaches are overcome. Experimental results, reported in Section 6, are encouraging both from the accuracy and from the computational points of view.

2 Passing from the Spectrum to the Composition or the Sequence

The MS/MS spectrum contains information about the structure but does not have any direct reference to the components of the polymer, being a mere succession of *peaks* corresponding to different molecular weights. The intensity of each peak is proportional to the number of molecules having that weight in the sample under analysis. A typical example is observable in Figure 1. Further processing is then requested, and usually performed, for peptides, as follows.

An initial *peak selection* phase is needed. This is generally done by removing all peaks below a certain intensity, since they are too noise-affected to be considered significant, and by assuming that all other peaks are informative. After this phase, the higher molecular weight among those corresponding to informative peaks is the one of the full peptidic complex under analysis, whereas the others correspond to its fragments. However, for each fragment peak, we

neither known the type of fragment which originated it (it could be one of the standard types a, b, c, x, y, z, or any other non-standard type) nor the number of electric charges that this fragment retained.

Now, some analysis techniques search for specific weight patterns in the spectrum and check them against similar patterns available from databases (e.g. [13]). However, when our compound is not in the databases (which may well happen: for many types of compounds, databases are not likely to be complete in the near future) or when the compound differs from the standard known form (protein sequences, for instance, often undergo modifications) alternative methods are required and direct identification is needed.

Direct identification, however, is not immediate. Moreover, the information contained in the spectrum may be insufficient for that. We will say that a combination or a sequence of components is *compatible* with a given spectrum if every informative peak in the spectrum admits an interpretation as a standard fragment of that combination or sequence. Often, however, there exists more than one composition or sequence which is perfectly compatible with a given spectrum. This means that the spectrum does not contain enough information to determine uniquely the composition or the sequence, and so there are more possibilities. Consider, for instance, the case of an incomplete fragmentation: if a part of a polymer never did break in the analysis, no detailed information on the inner structure of that part can be achieved. In this case, all the possible compositions or sequences compatible with the spectrum should be found, so as to guarantee accurate and objective character of the analysis. Sometimes it may also happen that a spectrum contains one or more peaks which have been selected as informative, but are instead due for instance to noise, non-standard fragmentation, spurious components. They are therefore not interpretable as standard fragments, so it may be the case that not even a composition or sequence exists that is compatible with the given spectrum. In this case, the best we can do is to be compatible with all but a number μ of peaks, and we will speak of μ-compatibility. This number of uninterpreted peaks is called the *mismatch* number μ.

An analysis procedure is therefore needed for passing from the spectrum to the composition or the sequence. These problems can be tackled by means of various solution approaches, each of them working on an abstract model of the problem. In order to analyze their characteristics, we distinguish between:

- The compositions and sequences that are compatible with the given spectrum but are not given and are the sets that one would like to find.
- The compositions and sequences that are given as outcome of the analysis procedure. These two sets may coincide with the former ones, depending on the quality of the adopted solution approach.

We will call *resolvents* of the spectrum the first two sets, while *results* of the procedure the latter ones. A solution approach is said to be *complete* if it guarantees finding as results all the possible resolvents of the spectrum; *incomplete* when such guarantee cannot be given, and therefore a part of the possible resolvents

may be neglected. This may also mean finding, in some cases, no resolvents at all. Moreover, a solution approach is said to be *exact* if it guarantees that every result given by the analysis is perfectly compatible with the given spectrum; *approximate* when this cannot be guaranteed and therefore the results are nearly compatible. Note the concept of an approximate result is more general and less precise than that of a μ-compatible solution. Nevertheless, due to the stochastic aspects involved in the fragmentation process, these approximate results may sometimes be probable solutions. As a matter of fact, complete and exact methods generally require larger computational times than incomplete or approximate ones (see e.g. [15]).

3 Related Work

For that which concerns direct peptide sequencing, known as *de novo* sequencing, some analysis procedures have been developed and implemented in a number of software systems, e.g. DeNovoX [17], Mass Seq[18], Peaks[19], Spectrum Mill[20]. Each of these procedures is essentially based on one of the following two approaches.

The first approach consists in searching the spectrum for couples of fragments belonging to the same standard type and differing by just one amino acid. That amino acid is therefore identified in the sequence. The whole sequence can be obtained in this manner when the spectrum contains a complete series of fragments. This is often unlikely since the fragmentation process is stochastic. Even though peptides tend to break at the conjunction of amino acids, they usually do not break at every conjunction of an amino acid, and furthermore, such cleavages may be of any of the mentioned different types. Further, if the collision energy is increased, the peptide produces more fragments but may break also at locations that are not the conjunction of amino acids, producing some non-standard fragments. Therefore, even though it is exact, the above approach should be classified as incomplete.

The second approach consists in iteratively generating a large number of virtual sequences by using a Monte Carlo method [4] and evaluating the match of the corresponding (theoretical) mass patterns with the (actual) mass pattern of the spectrum under investigation. Therefore, sequences producing a spectrum similar to the one under analysis can be obtained, but no completeness can be guaranteed. The number of possible peptides is very large: for example, there are $20^{12} \approx 10^{15}$ possible peptides composed of 12 amino acids, choosing them among 20 possible amino acid types. If one could generate and check 10^5 sequences per second, which for current computer seems optimistic, after 10^4 seconds of computation (almost 3 hours) 10^9 sequences would have been tried. This constitutes only a relatively small part of the set of all possible sequences (one every 10^6 in the example). Therefore, only a negligible portion of the solution space would have been explored, and there could be many sequences producing a spectrum much more similar to the one under analysis that have not been considered. Even by protracting the search or increasing the search

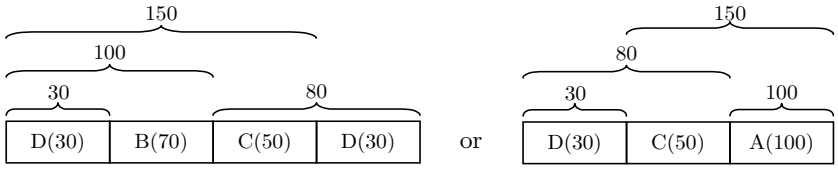
speed, when the number of generated sequences becomes near to the number of possible ones, no guarantee of repeating the same sequences can be given. This would require memorizing all the tested ones, and checking all of them after the generation of each new one, which is clearly impossible to accomplish in reasonable times with current computer technology [8]; or generating them in some ordered manner, and not by means of Monte Carlo methods. Finally, the similarity of spectra must be evaluated by choosing some similarity criterion. This has the consequence that the approach becomes an approximate one. The above described de novo techniques suffer therefore from considerable structural limitations.

Due to its combinatorial nature, the problem has also been recently approached by means of discrete mathematics. Specifically for the peptide sequencing problem, two approaches have been proposed. The first one is the graph theoretical construction proposed in [6], which evolved into the dynamic programming algorithms proposed in [1, 5]. The second one is the branching-based algorithm proposed in [3]. The first approach has the advantage of requiring a computational time for finding each solution which is polynomial, hence tractable [8], when imposing some limitations to the problem, namely no multi-charged fragments can appear in the spectrum, and only peaks corresponding to a set of fragment types which is "simple" [1] (e.g. only a-ions, b-ions and y-ions) can appear in the spectrum. When overriding such limitations, polynomial time cannot be guaranteed, and in any case the procedure cannot work with a spectrum in which all types of fragments and of charges may appear. The problem is NP-complete in the general case [1]. The second approach, on the other hand, has no structural limitations regarding types of fragments and charges, and performs a complete search. It requires, however, a heavier computational load. Therefore, peptides above a certain dimension cannot be satisfactory considered.

4 The Proposed Mathematical Models

A series of peaks must initially be selected as informative and extracted from the MS/MS spectrum. This is generally done by removing all peaks below a certain intensity, usually about 10% of the maximum intensity peak. Note that a different and more sophisticated peak selection may be performed (and this remains an open problem) without changing what follows. Such selected peaks give us the molecular weight of the full ionized compound (the peptide) and those of its ionized fragments. Generally speaking, not all fragments are ionized, so the presence in the spectrum of a fragment does not guarantee the presence of its complement. On the other hand, the molecular weights of the possible components (the amino acids) are known. By using these two types of information, the structure is investigated as highlighted in the following example.

Example 2.1. Given a compound of weight 180, fragment peaks of weights 150, 100, 80, 30, and a list of possible components A, B, C, D, of weights 100, 70, 50, 30, imagine that the weight of a compound is simply the sum of the weights of its components and that there are only two types of fragments: those containing the beginning and those containing the end of the compound. These simplifications are only used for this introductory example and do not generally hold for amino acids and peptides. This case has a nonunique solution. The compound can in fact be:



The possible combinations of components compatible with the spectrum are two: (1 of B, 1 of C, 2 of D), or (1 of A, 1 of C, 1 of D). The possible sequences of components compatible with the spectrum are four: (D-B-C-D), (D-C-B-D), (D-C-A), (A-C-D).

In order to simplify the exposition, we assume all the weight values to be rounded to integer numbers. Experimental practice in peptide sequencing shows that such precision is often adequate. There are of course no theoretical impediments to the use of greater precision in the described procedure, since the combinatorial aspects remains the same. Consider for instance Example 2.1. with non-integer input weights 181.8, 151.5, 101, 80.8, 30.3 and component weights 101, 70.7, 50.5, 30.3.

Peptides usually fragment due to the cleavage of one of its internal bonds. Any fragment containing the N-terminus is called N-terminal, while any one containing the C-terminus is called C-terminal. The cleavage of more then one bond in such a way of producing non-terminal fragments is rare and therefore not considered here. The six mentioned standard types of fragments (a, b, c, x, y, z) are described in Figure 2. When the broken bond is the one between the CO group of one amino acid and the NH group of another amino acid, the N-terminal part is called fragment of type b, or simply b-ion, and the C-terminal part is called y-ion. If the broken bond is the one between the central C of one amino acid and the CO group of the same amino acid, the N-terminal part is called a-ion, and the C-terminal part is called x-ion. If the broken bond is the one between the NH group of one amino acid and the central C of the same amino acid, the N-terminal part is called c-ion, and the C-terminal part is called z-ion.

Each of these standard fragments has a weight which is not simply the sum of the weights of its components, but which depends on them according to the following criteria: a b-ion weighs 1 plus the sum of the weights of its component amino acids, each of which decreased by 18; an a-ion weighs as the corresponding

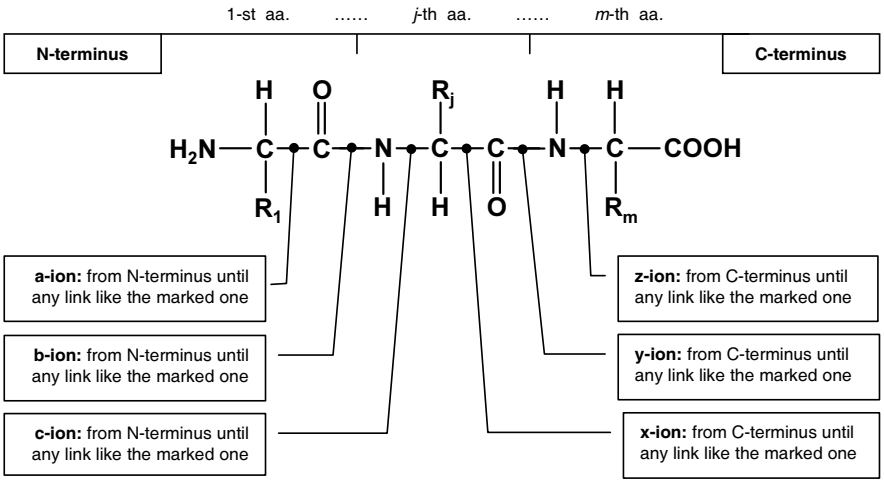


Figure 2: Standard peptide fragmentation.

b-ion minus 28; a c-ion weighs as the corresponding b-ion plus 17; a y-ion weighs 19 plus the sum of the weights of its component amino acids, each of which decreased by 18; an x-ion weighs as the corresponding y-ion plus 26; a z-ion, finally, weighs as the corresponding y-ion minus 16. Moreover, when a fragment retains more than one ionizing electric charge, the observed weight is a fraction of the actual ion weight. For instance, by denoting with w the weight of a generic mono-charged fragment, the corresponding double charged fragment has an actual weight of $w + 1$ due to the additional ionizing particle, and its observed weight is $(w + 1)/2$.

Non-standard fragmentation generally consists of additional fragmentation (losses of small neutral molecules such as water, ammonia, carbon dioxide, carbon monoxide, or breaking of a lateral chain), which lowers the observed weight of a fragment. Weight rules are, in this cases, quite complex and difficult to predict. Finally, noise peaks or peaks due to some spurious component may sometimes be present. For these reasons, depending on each specific analysis, the number of possibly non-standard peaks must be evaluated.

We now formalize the previously described problems as mathematical models, defined through a set of decision variables, having values inside given domains, and through a set of constraints.

- Denote by n the number of possible component amino acids (e.g. 20);
- by $A = \{1, 2, \ldots, n\}$ the set of indices corresponding to such amino acids, for instance in increasing weight order;
- by Maa_i, with $i \in A$, the weight of each amino acid (as usual in biochemistry);
- by MH^+ the weight of the precursor ion (as usual in biochemistry);

- by m the (unknown) number of amino acidic molecules contained in the analyzed peptide;
- by $m_{\max}$ and $m_{\min}$ respectively the maximum and the minimum possible value for the above m. Note that, if not obtained from other sources, they can be computed as

$$m_{\max} = \lfloor \mathrm{MH}^+ / \mathrm{Maa}_1 \rfloor \quad \text{and} \quad m_{\min} = \lceil \mathrm{MH}^+ / \mathrm{Maa}_n \rceil;$$

- by $B = \{1, 2, \ldots, m\}$ the set of indices corresponding to the amino acidic molecules contained in the analyzed peptide orderer from the N-terminal to the C-terminal, or in other words, the set of the indices of the positions within the analyzed peptide;
- by $t+1$ the number of peaks selected as informative in the spectrum, so that the number of informative peaks corresponding to fragments is t;
- by $F = \{1, 2, \ldots, t\}$ the set of indices corresponding to such fragment peaks;
- by p^l the weight of the l-th fragment peak;
- by H the set of the possible types of fragments, i.e. $\{a, b, c, x, y, z\}$;
- by μ the number of possibly non-standard peaks among the selected fragment peaks, i.e. the above mentioned mismatch number.

In the case of the problem of finding the composition, the following set of variables is used:

$$x_i = \text{ number of molecules of the } i^{\text{th}} \text{ amino acid in the peptide, } \ i \in A.$$

Such a number of molecules has to be a positive integer between a minimum $x_{i\,\min}$ and a maximum $x_{i\,\max}$ value. This also allows one to account for the known presence or absence of some amino acid. Note that, if not obtained from other sources, they can be computed as

$$x_{i\,\max} = \lfloor \mathrm{MH}^+ / \mathrm{Maa}_i \rfloor \quad \text{and} \quad x_{i\,\min} = 0.$$

Therefore, each one of the x_i variables has its admissible domain:

$$x_i \in \mathbf{Z}_+ \cap [x_{i\,\min}, x_{i\,\max}].$$

Moreover, in order to consider that each fragment l may contain some of the amino acid molecules of the full peptide, an additional set of variables is used:

$$y_i^l = \text{number of molec. of the } i^{\text{th}} \text{ amino acid in } l^{\text{th}} \text{ fragment, } \ i \in A, l \in F.$$

The y^l variables are positive integers such that $y_i^l \leq x_i$ for each i. We briefly denote this relationship between the two vectors by $y^l \leq x$.

Since the number of electric charges retained by the l-th fragment modifies its observed weight p^l, variables for expressing such number of charges are needed:

$$e^l = \text{number of electric charges in } l^{\text{th}} \text{ fragment, } \ l \in F.$$

They have a value between 1 and a maximum value $e_{\max}$, i.e. $e^l \in \mathbf{Z}_+ \cap [1, e_{\max}]$.

In the case of the problem of finding the sequences, the following set of binary variables is used:

$$w_{ij} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ amino acid is in position } j^{\text{th}} \text{ of peptide, } i \in A, j \in B \\ 0 & \text{otherwise.} \end{cases}$$

In order to consider that each fragment l corresponds to a subsequence $S^l = \{u^l, u^l + 1, \ldots, v^l\}$ of B such that either $u^l = 1$ or $v^l = m$ but not both, two additional sets of variables are used:

$u^l =$ index of the position of first amino acid of subsequence $S^l,\ l \in F$;
$v^l =$ index of the position of last amino acid of subsequence $S^l,\ l \in F$.

Clearly, $u^l \in B$ and $v^l \in B$. For N-terminal fragments $u^l = 1$ and $v^l < m$, for C-terminal ones $u^l > 1$ and $v^l = m$.

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

- Denote by k and k^+ two constants respectively representing the weight that each amino acid loses when it is inside a peptidic chain and the weight that must be added for obtaining the weight of the precursor ion. Their values, when using integer weights, are respectively 18 and 19.
- Denote, finally, by $k_a, k_b, k_c, k_x, k_y, k_z$ the constants representing the weights that must be added for obtaining the weight of the different types of standard fragments. Their values, when using integer weights, are respectively $\{-27, 1, 18, 45, 19, 3\}$.

In the case of the problem of finding the combinations of amino acids, the constraint securing compatibility with the overall weight of the compound is:

$$k^+ + \sum_{i \in A} [x_i(\text{Maa}_i - k)] = \text{MH}^+. \tag{1}$$

The above can be interpreted as each amino acid losing a weight of k when inside the peptidic chain, except for the N-terminus which loses only 17 and the C-terminus which loses only 1. The ionized peptide gains an additional weight of 1 due to the effect of the ionizing particle.

The constraints securing compatibility with the peaks selected in the spectrum are of the type $\exists\, y^l \leq x$ such that, for each $l \in F$, there is a value $\hat{k} \in \{k_a, k_b, k_c, k_x, k_y, k_z\}$ such that:

$$\hat{k} + \sum_{i \in A} [y_i^l(\text{Maa}_i - k)] = (p^l * e^l) - (e^l - 1). \tag{2}$$

We call these *fragment constraints* for the composition problem. They mean that each peak of observed weight p^l is either an a-ion, a b-ion, a c-ion, an x-ion, a y-ion, or a z-ion that retained e^l electric charges. It therefore had a real

weight of $(p^l-1)e^l+1$. In order to express that $\hat{k} \in \{k_a, k_b, k_c, k_x, k_y, k_z\}$ we need another set of binary variables:

$$z_h^l = \begin{cases} 1 & \text{if the } l^{\text{th}} \text{ fragment is of type } h,\ l \in F, h \in H \\ 0 & \text{otherwise.} \end{cases}$$

The above fragment constraint for the $l-$th peak becomes therefore

$$\begin{array}{l} \displaystyle\sum_{h\in H} k_h z_h^l + \sum_{i\in A} [y_i^l(\text{Maa}_i - k)] = (p^l * e^l) - (e^l - 1) \\ \displaystyle\sum_{h\in H} z_h^l = 1. \end{array}$$

Altogether, the set of constraints representing the composition problem is

$$\begin{cases} k^+ + \displaystyle\sum_{i\in A} [x_i(\text{Maa}_i - k)] = \text{MH}^+ & \\ \displaystyle\sum_{h\in H} k_h z_h^l + \sum_{i\in A} [y_i^l(\text{Maa}_i - k)] = (p^l-1)e^l + 1 & \forall l \in F \\ \displaystyle\sum_{h\in H} z_h^l = 1 & \forall l \in F \\ x_i \le \text{MH}^+/\text{Maa}_i & \forall i \in A \\ y_i^l \le x_i & \forall i \in A, l \in F \\ e^l \le e_{\max} & \forall l \in F \\ z_h^l \in \{0,1\} & \forall h \in H, l \in F \\ x_i, y_i^l, e^l \in \mathbf{Z}_+ & \forall i \in A, l \in F. \end{cases} \tag{3}$$

In the case of the problem of finding the sequences of amino acids, the constraint securing compatibility with the overall weight of the compound is instead:

$$k^+ + \sum_{i\in A}\sum_{j\in B} [w_{ij}(\text{Maa}_i - k)] = \text{MH}^+, \tag{4}$$

while the constraints securing compatibility with the weight of the various types of fragments introduced above are of the type $\exists\, m, w, u^l, v^l$ such that, for all $l \in F$ there is a value $\hat{k} \in \{k_a, k_b, k_c\}$ and a value $\hat{\hat{k}} \in \{k_x, k_y, k_z\}$ such that:

$$\begin{array}{c} \hat{k} + \displaystyle\sum_{i\in A}\sum_{j=1}^{v^l} [w_{ij}(\text{Maa}_i - k)] = (p^l * e^l) - (e^l - 1) \\ \text{or} \\ \hat{\hat{k}} + \displaystyle\sum_{i\in A}\sum_{j=u^l}^{m} [w_{ij}(\text{Maa}_i - k)] = (p^l * e^l) - (e^l - 1). \end{array} \tag{5}$$

We call these *fragment constraints* for the sequencing problem. They mean that each peak of observed weight p^l is either an a-ion, a b-ion, or a c-ion if considering an N-terminal subsequence, or either an x-ion, a y-ion, or a z-ion if

considering a C-terminal sequence, that retained e^l electric charges. It therefore had a real weight of $(p^l-1)e^l+1$. By using the above introduced z_h^l variables, the above fragment constraints for the $l-$th peak become the following disjuntive constraints

$$\begin{array}{l} \displaystyle\sum_{h\in\{a,b,c\}} k_h z_h^l + \sum_{i\in A}\sum_{j=1}^{v^l} [w_{ij}(\mathrm{Maa}_i - k)] = (p^l * e^l) - (e^l - 1) \\ \displaystyle\sum_{h\in\{a,b,c\}} z_h^l = 1 \end{array}$$

$$\bigvee$$

$$\begin{array}{l} \displaystyle\sum_{h\in\{x,y,z\}} k_h z_h^l + \sum_{i\in A}\sum_{j=u^l}^{m} [w_{ij}(\mathrm{Maa}_i - k)] = (p^l * e^l) - (e^l - 1) \\ \displaystyle\sum_{h\in\{x,y,z\}} z_h^l = 1. \end{array}$$

Finally, constraints imposing that for each position of the peptide there is exactly one amino acid are used

$$\sum_{i\in A} w_{ij} = 1 \qquad\qquad \forall j \in B. \tag{6}$$

Altogether, the set of constraints representing the sequencing problem is:

$$\left\{ \begin{array}{ll} k^+ + \displaystyle\sum_{i\in A}\sum_{j\in B} [w_{ij}(\mathrm{Maa}_i - k)] = \mathrm{MH}^+ & \\ \left(\begin{array}{l} \displaystyle\sum_{h\in\{a,b,c\}} k_h z_h^l + \sum_{i\in A}\sum_{j=1}^{v^l} [w_{ij}(\mathrm{Maa}_i - k)] = (p^l * e^l) - (e^l - 1) \\ \displaystyle\sum_{h\in\{a,b,c\}} z_h^l = 1 \\ \qquad\qquad \bigvee \\ \displaystyle\sum_{h\in\{x,y,z\}} k_h z_h^l + \sum_{i\in A}\sum_{j=u^l}^{m} [w_{ij}(\mathrm{Maa}_i - k)] = (p^l * e^l) - (e^l - 1) \\ \displaystyle\sum_{h\in\{x,y,z\}} z_h^l = 1 \end{array} \right) & \forall l \in F \\ \mathrm{MH}^+/\mathrm{Maa}_n \le m \le \mathrm{MH}^+/\mathrm{Maa}_1 & \\ 2 \le v^l \le m & \forall l \in F \\ u^l \le v^l - 1 & \forall l \in F \\ e^l \le e_{\max} & \forall l \in F \\ w_{ij} \in \{0,1\} & \forall i \in A,\ j \in B \\ z_h^l \in \{0,1\} & \forall h \in H, l \in F \\ m, u^l, v^l, e^l \in \mathbf{Z}_+ & \forall l \in F. \end{array} \right. \tag{7}$$

We are interested in identifying all solutions to the above sets of constraints, written for all the peaks selected in the spectrum, except at most the number of possibly non interpretable peaks μ. In the case of composition, this means all the vectors $x = (x_1, x_2, \ldots, x_n)$ which may not verify at most μ fragment constraints of type (2). In the case of sequencing, it means all the vectors $w = (w_{11}, w_{12}, \ldots, w_{1m}, \ldots, w_{n1}, w_{n2}, \ldots, w_{nm})$ which may not verify at most μ fragment constraints of type (5).

The problem is therefore a feasibility problem: all the above vectors x and w provide an interpretation for all peaks, except at most μ of them. They are therefore all the solutions μ-compatible with the given spectrum, and, for any fixed μ, there is no general objective function to choose among them. The presence, in some cases, of many solutions, simply means that the spectrum does not contain enough information to obtain a more precise solution. Objective functions based on the intensities of interpreted peaks, as proposed in [1], may also be considered, but they do not appear to be universally valid since, once a peak has been selected as informative, its relevance for the analysis is not directly proportional to its intensity.

Since some molecular weights cannot correspond to any amino acidic combination, in some cases we are able to determine in advance that a peak cannot be originated by certain types of fragment. This additional information is particularly useful and should be introduced in our model by excluding the possibility of checking those types of fragment for that peak in constraints (2) or (5).

Example 2.2. Consider a MS/MS spectrum producing the following data:

$$\mathrm{MH}^+ = 327,\ F = \{p^1, p^2, p^3\},\ p^1 = 155,\ p^2 = 76,\ p^3 = 58,\ \mu = 0.$$

In the case of the sequencing problem, we have 1 constraint of type (3), 18 possible constraints (if expanding the disjunctions) of type (4) and 4 constraints of type (5). The problem is therefore small. The solution is

N-Gly-Pro-Pro-Gly-OH,

whose weight is the sum of the weights of two Gly (75 each) and two Pro (115 each), every one decreased by 18, plus 19: 57+57+97+97+19=327. Fragment 155 can be interpreted as the b-ion H-Gly-Pro, whose weight is 57+97+1=155; fragment 76: can be interpreted as the y-ion Gly-OH, whose weight is 57+19=76; fragment 58: can be interpreted as the b-ion H-Gly, whose weight is 57+1=58.

5 Solution Algorithms

As one can see, the above described problems quickly become of very large dimension when increasing MH^+ and n. In order to give an illustration, there are more than 10^{10} possible different amino acidic sequences having the weight MH^+=1000. The presence of disjunctions may be represented in an integer linear programming problem by introducing additional binary variables. This,

however, would produce a substantial model, that is difficult to solve [8]. The need for generating all the compatible solutions (and not just one) does not suggest either the use of some type of search heuristic, since it would be difficult to determine whether all possible solutions have been found.

On the other hand, specialized *branching techniques* are in many cases efficient on similar problems. Examples are the cases of propositional satisfiability, see e.g. [9], or of several combinatorial optimization problems, see e.g. [12, 15]. An important feature is that they can work on mixed logical and mathematical expressions, so disjunction elimination, with consequent model explosion, is not necessary. Moreover, they can guarantee that all solutions are found when the search tree is completely explored. Therefore, search techniques based on branching are used directly on the sets of constraints (1, 2) and (4, 5, 6). The proposed solution approach is consequently complete and exact. Such techniques rely on systematic and recursive partitioning of the search space in regions easier to be explored. This is achieved by progressively *fixing* variables to values ($v_i \in D_i$ for the x_i variables, $v_{ij} \in D_{ij}$ for the w_{ij} variables) thus generating subproblems with progressively decreasing dimension. The search evolution may be represented as a *search tree*. Each non-leaf node of the search tree corresponds to a partial solution. In order to expedite the search tree exploration, branches that do not yield solutions are not to be explored. This is generally performed by verifying if the current branch of the search tree corresponds to a partial solution not respecting all the constraints.

In particular, we derive two basic branching and backtracking algorithms to search all combinations and sequences. Since they are very similar, we report the algorithm for the case of sequences by writing in square brackets the differences with respect to the case of the combinations. Note that in step 3 (TestFrag) a current solution S_c is said to *clash* with a peak p_l when S_c not only does not satisfy one of the constraints given by p_l but is such that keeping the values of the variables which are fixed in S_c would prevent all the constraints given by p_l to be satisfied, even by fixing in every possible way the unassigned variables. In this case backtrack should be immediately performed. On the contrary, backtracking performed only when a current partial solution explicitly violates all the constraints given by peak p_l would be too belated. In the fix and backtrack steps, values and variables must be chosen in a pre-ordered manner, to avoid cycling. We denote in curly brackets sets of values, such as $\{p_l\}$.

Algorithm Search Combinations [resp. **Sequences**]
Input: *compound data* MH^+, $\{p^l\}, \mu$, *components data* $A, \{\mathrm{Maa}_i\}$.
Output: *all vectors* x [resp. w] *compatible with input data.*

1. Fix: *choose an unassigned variable as the current variable* x_c [resp. $w_{cc'}$] *and fix it to a current value* $v_c \in D_c$, [resp. $v_{cc'} \in \{0, 1\}$], *if not possible goto* Backtrack

2. TestFull: *if the current solution* S_c *violates constraint* (1) [resp. (4) *or* (6)], *goto* Backtrack

3. TestFrag: *if S_c clashes with more than μ peaks in $\{p_l\}$*, Backtrack

4. Output: *if S_c is a complete solution, output it*

5. Termination: *if the search tree is completely explored, stop; otherwise goto* Backtrack

6. Backtrack: *fix x_c* [resp. $w_{cc'}$] *to the next value $\bar{v}_c \in D_c$* [resp. $\bar{v}_{cc'} \in \{0,1\}$] *if possible, otherwise release x_c* [resp. $w_{cc'}$], *choose another variable having at least one unused value in its domain as the new x_c* [resp. $w_{cc'}$], *fix it to that value, goto* TestFull

The above algorithm is computationally satisfactory for the case of the problem of finding the combinations of amino acids. Note also that such a problem has a field of interest more limited than the problem of finding the sequences (for instance, it should be solved in the case of *cyclic* peptides). In the case of the sequencing problem, on the contrary, the above algorithm is not computationally able to solve problems that are of practical interest (see Table 1) because it requires rapidly increasing computational times. One can observe that the described combinatorial constraints ((2) for the composition problem, (5) for the sequencing problem) are challenging to verify since each of them requires a large number of possible subsets of the current solution. Moreover, for a partial solution, it is rarely the case that similar constraints are definitely violated. So, especially at the first levels of the explored branching tree, the checking of such constraints is almost useless and overly time consuming. On the other hand, the constraints on the overall weight of the compound ((1) for the composition problem, (4) for the sequencing problem) and the other constraints not involving fragments are easier to verify.

Therefore, we propose a specialized algorithm for the sequencing problem. According to the same principles, an improved algorithm for the problem of finding the composition could be developed. The only constraints that are verified during the whole branching tree exploration are those of type (4) and (6), and possibly fragment constraints for which the type of fragment is known and which must certainly hold, regardless of μ (under special conditions such information may be available, as noted at the end of Section 4). We call these latter *firm* constraints. When a leaf of the search tree is reached, a complete solution $(v_{11}, v_{12}, \ldots, v_{1m}, \ldots, v_{n1}, v_{n2}, \ldots, v_{nm})$ is at hand. Such a solution is then checked to see if it satisfies the remaining fragment constraints (5). Clearly, if it verifies those obtained for all p_l, except at most μ of them, that is one of the feasible solutions of the problem. In the opposite case, that is not a solution, and we continue the search tree exploration by backtracking. The number of nodes of the search tree enumerated by this algorithm is greater than that of the basic algorithm, since it can only cut part of the branches that can be cut by the basic algorithm However, processing each node is faster. The approach of delaying the checking of a set of constraints is inspired by the well-known techniques of *delayed row generation* [2].

Algorithm Search Sequences Improved
Input: *compound data* MH^+, $\{p^l\}, \mu$, *components data* $A, \{\mathrm{Maa}_i\}$.
Output: *all vectors* w *compatible with input data.*

1. Fix: *choose an unassigned variable as the current variable* $w_{cc'}$ *and fix it to a current value* $v_{cc'} \in \{0, 1\}$, *if not possible goto* Backtrack
2. TestFull: *if the current solution* S_c *violates a constraint of type* (4) *or* (6) *or a firm fragment constraint*, *goto* Backtrack
3. TestFrag: *if* S_c *is a complete solution, test whether it satisfies all but* μ *constraints of type (5). If yes output it*
4. Termination: *if the search tree is completely explored, stop; otherwise goto* Backtrack
5. Backtrack: *fix* $w_{cc'}$ *to the opposite value* $\bar{v}_{cc'} \in \{0, 1\}$ *if possible, otherwise release* $w_{cc'}$, *choose another variable still having at least one unused value in its domain as the new* $w_{cc'}$, *fix it to that value, goto* TestFull

Since the number of amino acids in the sequence is not known in advance, the number of w_{ij} variables is not known during the search. This does not represent a problem, since variables are generated until a current partial solution violates constraint (4), and then, if needed, removed. Finally, we should remark that fragments corresponding to b-ions and y-ions are by far the most common. Therefore, their detection overrides the detection of the other types of fragments.

When imposing no limitations to the problem, the computational worst-case time complexity of the described algorithms is exponential (like similar algorithms based on branching and backtracking). This is usual since the above problems are NP-complete in general. The subproblem of finding only a single solution (either composition or sequence) is in P when having only a reduced set of types of monocharged fragments [1].

6 Experimental Results

The proposed algorithms were implemented in C++ and ran on a Pentium IV 1.7GHz PC with 1Mb RAM. An initial routine reads spectrometry data, checking whether or not each peak can be only a specific type of fragment. It then reads the list of possible components, which includes the usual 20 amino acids, their modified versions (obtained due to glycosylation, phosphorylation, acetylation, methylation, etc.) when their presence is suspected and any other spurious component whose presence is suspected. The procedure then searches for all possible solutions and reports them lexicographically ordered by the molecular weights of the components, modulo a permutation. Results of the analysis on MS/MS spectra are presented in Figure 3 and 4. Note that in the former case,

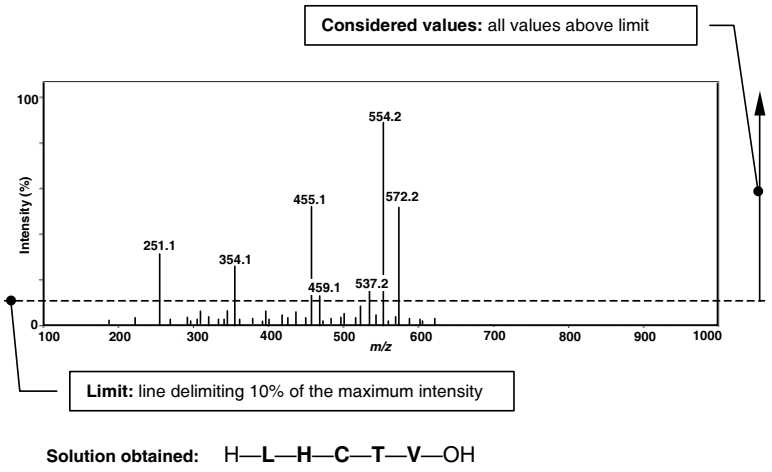


Figure 3: Sequencing of a medium size peptide.

the information contained in the spectrum is enough to determine a unique sequence. In the latter case, the spectrum does not contain enough information for the determination of a unique sequence, and therefore more then one solution is identified (in particular, four). These results have been confirmed by means of various biological exams (subsequent peptide analysis, or peptide synthesis, etc.)

From the computational perspective, we report in Table 1 the results obtained for the processing of several sets of data representing peptides of increasing dimension. In particular, we give the full weight of the peptide (MH^+), the number of considered peaks (# peaks), the acceptable mismatch number μ (Mism.), the number of solutions (# of sol.) of the sequencing problem (the number of combinations may be different, by definition of the problem), and computational times (in CPU seconds) required for solving the problem of finding the combinations (Comb.) and the sequences. Results for both the basic (Seq. 1) and the improved (Seq. 2) versions of the algorithm are shown. We set a time limit of 3600 sec., and report † when this is exceeded.

The elapsed time increases with the molecular weight of the processed peptide. This weight clearly causes a growth in the size of the problem. Moreover, the elapsed time increases with the number of considered peaks and with the value of acceptable mismatches. Consider for instance the three lines of Table 1 reporting the results for 851.3. They represent the same peptide submitted to MS/MS spectrometry under different conditions, the intuition being that the higher the number of peaks and the given value of acceptable mismatches, the more *noisy* the mass spectrometry analysis has been. The value of acceptable mismatches represents exactly this. Therefore, the results become less precise

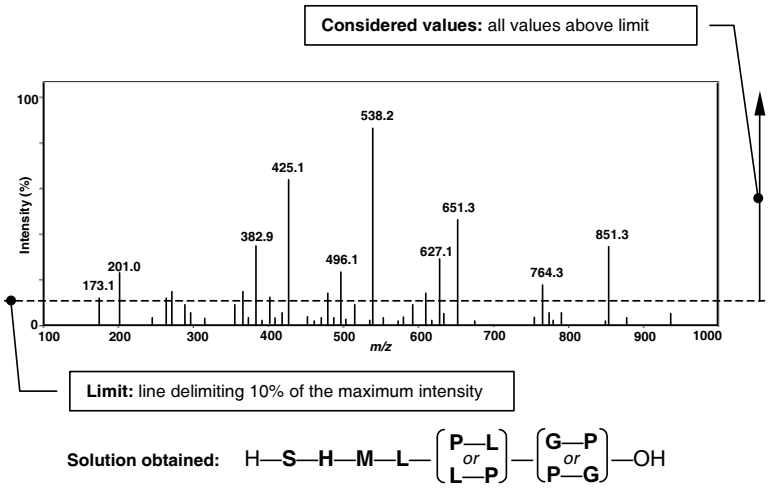


Figure 4: Sequencing of a moderately large size peptide.

and require more time. By comparing the different algorithms for the sequencing problem, the improved version is faster. In fact, even if the improved algorithm enumerates more nodes of the search tree, its node processing is faster, and therefore its overall behavior is preferable.

7 Conclusions

The determination of all possible combinations, or all possible sequences, of an unknown polymer submitted to MS/MS spectrometry analysis is a basic and relevant problem. Since the fragmentation processes occurring during such analysis often have complex and non-deterministic rules, known solution approaches typically suffer from limitations. The problem is approached here by developing a mathematical model of the fragmentation process. Such a model is exemplified by considering the case of peptides, but, with immediate modifications, may be used for generic polymeric compounds submitted to mass spectrometry. The proposed approach also allows one to tackle cases when the spectrum does not contain enough information for a univocal determination of the solution, or when the spectrum contains uninterpretable peaks originated by non-standard fragmentation, impurities or noise. Due to its combinatorial nature such models are computationally demanding to solve. Specialized algorithms based on branching techniques are proposed. Results are encouraging and validated.

Table 1: Computational behavior for different peptides in increasing weights.

Spectrum		Analysis		Time (in CPU secs.)		
MH^+	# peaks	Mism.	# of sol.	Comb.	Seq. 1	Seq. 2
424.0	3	0	1	0	0	0
535.1	5	0	3	1	1	1
571.0	2	0	12	2	4	1
572.4	4	0	1	1	5	1
635.2	6	0	8	8	123	1
669.0	5	0	12	6	148	1
700.5	6	0	2	5	2510	1
759.2	6	0	1	50	†	1
851.3	8	2	4	71	†	20
851.3	18	10	4	†	†	73
851.3	21	14	24	†	†	101
859.1	7	0	4	312	†	8
968.6	12	5	30	†	†	2280
993.5	10	1	104	74	†	32
1029.4	6	0	329	330	†	77

Acknowledgments

The author is grateful to Prof. G. Gianfranceschi and to Prof. G. Koch for stimulating discussions and for providing clarifications on the biological aspects of the problems, and to Prof. L. De Angelis and Mr. F. Giavarini for the mass spectrometry experimental analysis.

References

[1] Bafna V. and N. Edwards. On de novo interpretation of tandem mass spectra for peptide identification In *Annual Conference on Research in Computational Molecular Biology* (RECOMB03), 9-18 (2003).

[2] Bertsimas D. and J.N. Tsitsiklis. *Introduction to Linear Optimization*. (Athena Scientific, Belmont, Massachusetts, 1997).

[3] Bruni R., G. Gianfranceschi, and G. Koch. On Peptide De Novo Sequencing: a New Approach. *Journal of Peptide Science*, 11, 225-234 (2005).

[4] Casella G. and C.P. Robert. *Monte Carlo Statistical Methods*, (Springer, New York, 2006).

[5] Chen T., M.Y. Kao, M. Tepel, J. Rush, and G.M. Church. A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology*, 8(6), 571-583 (2001).

[6] Dancik V., T.A. Addona, K.R. Clauser, J.E. Vath, P.A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology* 6, 327-342 (1999).

[7] Edman P. A method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* 4, 283-293 (1950).

[8] Garey M.R. and D.S. Johnson. *Computers and Intractability: a Guide to the Theory of NP-Completeness*, (Freeman, New York, 1979).

[9] Gu J., P.W. Purdom, J. Franco, B.W. Wah. Algorithms for the Satisfiability (SAT) Problem: A Survey. *DIMACS Series in Discrete Mathematics*, 35, 19-151 (1997).

[10] Heinrikson R.L. The Edman Degradation in Protein Sequence Analysis. In: Lo, T.B., T.Y. Liu, C.H. Li, (editors) *Biochemical and Biophysical Studies of Proteins and Nucleic Acids*, (Elsevier, New York, 1984) 285-302.

[11] Holm L. and C. Sander. Protein Structure Comparison by Alignment of Distance Matrices. *Journal of Molecular Biology.* 233, 123-138 (1993).

[12] Hooker J.N. *Logic Based Methods for Optimization*, (Wiley, New York, 2000).

[13] Johnson R.S. and J.A. Taylor. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Methods Mol. Biol.* 146, 41-61 (2000).

[14] Montaudo G. and R.P. Lattimer (editors). *Mass Spectrometry of Polymers*. (CRC Press, 2001).

[15] Nemhauser G.L., L.A. Wolsey. *Integer and Combinatorial Optimization*, (Wiley, New York, 1988).

[16] Siuzdak G. *Mass Spectrometry for Biotechnology*, (Academic Press, New York, 1996).

[17] Software system DeNovoX. ThermoFinnigan Corp. (http://www.thermo.com).

[18] Software system Mass Seq. Micromass Ltd. (http://www.micromass.co.uk).

[19] Software system PEAKS. Bioinformatics Solutions Inc. (http://www.bioinformaticssolutions.com).

[20] Software system Spectrum Mill. Agilent Technologies Inc. (http://www.agilent.com).

[21] Taylor J.A. and R.S. Johnson. Implementation and Uses of Automated De Novo Peptide Sequencing by Tandem Mass Spectrometry. *Analytical Chemistry* 73(11), 2594-2604 (2001).