# A Logic-Based Approach to Polymer Sequence Analysis[*]

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

September 5, 2009

**Abstract**

Polymers are compounds formed by the joining of smaller, often repeating, units linked by covalent bonds. The analysis of their sequence is a fundamental issue in many areas of chemistry, medicine and biology. Nowadays, the prevalent approach to this problem consists in using a mass spectrometry analysis that gives information about the molecular weights of the polymer and of its fragments. This information should be used in order to obtain the sequence. This is however a difficult mathematical problem, and several approaches have been proposed for it. In particular, a promising one is based on a propositional logic modeling of the problem. This paper presents conceptual improvements in this approach, principally the off-line computation of a database that substantially speeds-up the sequencing operations. This is obtained by finding a correspondence between sequences and natural numbers, so that all sequences up to a certain molecular weight can be implicitly considered in the above database, and explicitly computed only when needed. Results on real-world problems show the effectiveness of this approach.

**Keywords:** Mass Spectrometry, Peptide or Polymer Analysis, Propositional Logic Applications

## 1 Introduction

Polymeric compounds are formed by the joining of smaller units, here generically called *components*, linked by covalent bonds. The determination of the *sequence* of polymeric compounds is one of the most important and frequent issues in many areas of chemistry, medicine and biology, as well as in several

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

other applicative fields. It consists in finding which are all the components constituting the polymer under analysis, and the relative position of each of them. 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, and are usually constituted by a single sequence of components called *amino acids*.

Nowadays, a widely used and well established approach to sequence analysis consists in the use of mass spectrometry (e.g. [18, 21, 19]). Such technique can provide the absolute molecular weight distribution of a number of molecules in the form of a *spectrum*: for each molecular weight, the amount of material having that molecular weight produces a *peak* having a certain *intensity*. The study of the weight pattern in the spectrum can be used for understanding the structure of such molecules, expecially when using the mass spectrometry/mass spectrometry methodology (also known as MS/MS, or tandem mass, e.g. [27]). This procedure works as follows. After the first mass analysis, some molecules of the protonated polymer under analysis, called *precursor ion*, are selected and collided with other non reactive elements. This interaction leads to the fragmentation of many of such molecules, and the collision-generated decomposition products undergo a second mass analysis. Therefore, such analysis provides the absolute molecular weight of the full precursor ion, as well as those of the various ionized fragments obtained from that precursor ion. Non ionized fragments, on the contrary, do not appear in the spectrum. Such experiments may be performed by using several instrumental configurations, mainly triple quadrupole (QQQ), quadrupole time-of-flight (Q-TOF) and ion trap devices [19].

Since the weights of the possible components are known, and rules for determining the weights of sequences of known composition are available, one wants to use the MS/MS information in order to determine the unknown sequence of a polymer. This is however a difficult mathematical problem, as explained in detail in Section 2. Note that the presence of fragments constitutes the only source of information about the inner structure of the polymer under analysis: in absence of fragmentation, the inner structure would be unknown. Several approaches to this problem have been proposed, as reported in Section 3. In particular, a promising approach [4] is based on a propositional logic modeling (see e.g. [11, 17, 29]) of the problem, as explained in Sections 4 and 5. It can be shown that all and only the possible outcomes of a sequence analysis can be obtained by finding all models of a propositional logic formula. This paper presents conceptual improvements in this approach, principally the off-line computation of the so-called weights database, that substantially speeds-up the sequencing operations, as described in Section 6. This is obtained by finding a correspondence between sequences and natural numbers, so that all sequences up to a certain molecular weight can be implicitly considered in the above database, and explicitly computed only when needed. The procedure is exemplified by considering the case of peptides, but may be used for generic polymeric compounds submitted to mass spectrometry. Results on real-world problems, shown in Section 7, demonstrate the effectiveness of this approach.

## 2 From the Spectrum to the Sequence

The MS/MS spectrum contains our 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.

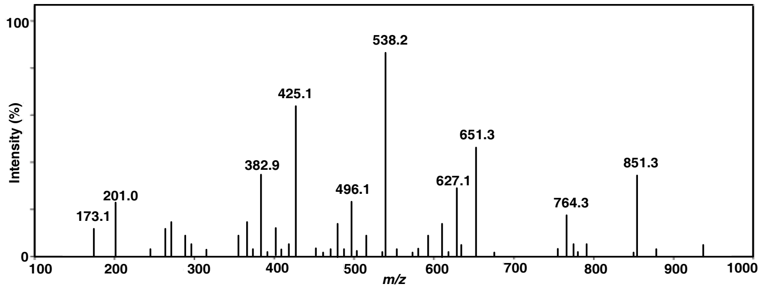


Figure 1: A MS/MS spectrum generated by collision-induced dissociation.

An initial *peak selection* phase is needed. This is generally done by removing all peaks below a certain intensity, since they are too noise-prone to be considered significant, and by considering informative all other peaks. After this phase, the higher molecular weight among informative peaks is the one of the full polymer under analysis, whereas the others correspond to its fragments. Though fragmentation is a stochastic process, some rules may be traced. The most abundant fragments are generally given by the cleavage of the weakest molecular bonds. Therefore, some types of fragments, called *standard* fragments, are more common than others, and should more likely correspond to the peaks selected as informative in the spectrum. In the case of peptides, for instance, there are six different types of standard fragments, called a, b, c, x, y, z. Fragments appear in the spectrum when ionized by retaining one or more electrical charges. Unfortunately, when analyzing each of such fragment peaks, we neither known the type of fragment which originated it (it could be any of the standard types, or also a 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 a databases of compounds (e.g. [16]). However, when our polymer is not in the databases (which may very well happen) or when the polymer differs from the standard known form (protein sequences, for instance, often undergo modifications) a constructive identification is required. Constructive identification, however, is not immediate, and, moreover, the information contained in the spectrum may be insufficient for that.

**Definition 2.1** We will say that 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 sequence.

Often, however, there exists more than one sequence which is perfectly compatible with a given spectrum. This means that the spectrum does not contain enough information to determine uniquely 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 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 sequence exists which is compatible with the given spectrum. In this case, the best we can do, informally speaking, is being compatible with as many peaks as it is possible.

**Definition 2.2** We will say that a sequence of components is *$\mu$-compatible* with a given spectrum if every informative peak in the spectrum, except a number $\mu$ of them, admits an interpretation as a standard fragment of that sequence. This number of uninterpreted peaks will be called the *mismatch* number $\mu$.

In order to analyze the features of the various approaches to the problem of passing from the spectrum to the sequence, we need to define the following sets.

**Definition 2.3** The *resolvents* of a spectrum are all the sequences which are compatible with the that spectrum (but are not given: are those that should be found).

**Definition 2.4** The *results* of a procedure are all the sequences which are given as outcome of the analysis procedure.

The above two sets may coincide or not, depending on the quality of the adopted solution approach.

**Definition 2.5** 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 could mean finding, in some cases, no resolvents at all.

**Definition 2.6** 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 given are only near-compatible, according to some nearness criterion.

Note that this concept of approximate results is more general and less precise than that of $\mu$-compatible solution. Nevertheless, due to the stochastic aspects involved in the fragmentation process, these approximate results may sometimes be probable solutions.

# 3 Related Work

For that which concerns constructive peptide sequencing, known as *de novo* sequencing, some analysis procedures have been developed and implemented in a number of software systems, e.g. DeNovoX [22], Mass Seq[23], Peaks[24], Spectrum Mill[25]). Each of such procedures is essentially based on one of the following two approaches.

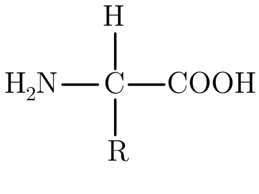
The first one consists in searching the spectrum for continuous series of fragments belonging to the same standard type and differing by just one amino acid, which is therefore identified. The whole sequence can be obtained in this manner when the spectrum contains a complete series of fragments. This,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

The second approach consists in iteratively generating, by using Monte Carlo methods [7], a large number of virtual sequences, 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 in fact very large: just for example, the possible peptides composed of 12 amino acids, choosing them among 20 possible amino acid types, are $20^{12} \approx 10^{15}$. So, even hypothesizing of generating and checking $10^5$ sequences per second, which for nowadays computer seems quite optimistic, after $10^4$ seconds of computation (almost 3 hours) only $10^9$ sequences would have been tried, which means a relatively small part of the possible ones (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. And, 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 do in reasonable times for nowadays computer technology [14]; 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, with the consequence that the approach becomes an approximate one. The above described analysis techniques suffer therefore form 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, there have been on the one hand, the graph theoretical construction proposed in [13], which evolved into the dynamic programming algorithms proposed in [10, 2], and, on the other hand, the branching-based algorithm proposed in [5], which evolved into the propositional logic modeling proposed in [4]. The first approach has the advantage of requiring a computational time for finding each solution which is polynomial, hence tractable [14], 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" [2] (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 in the general case is however NP-complete [2]. 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, so computational improvements would be useful for it.

# 4 A Mathematical View of the Fragmentation Process

When a polymer undergoes a MS/MS analysis, the occurring fragmentation process gives an essential support to the sequencing. We now analyze in detail peptide fragmentation as an explanatory example of any generic fragmentation process. Similar analyses may be performed of course also for other categories of polymers. Peptides basically are single sequences of building-blocks called *amino acids*. Each amino acid molecule has the following general chemical structure.



There is a large number of possible amino acids, differing in the internal chemical structure of the radical R, and, therefore, for their functional characteristics and their molecular weights. The most commonly considered ones generally include those reported in Table 1. Moreover, each amino acid may also present one of the many possible modifications, such as phosphorylation, acetylation, methylation, etc. This would produce alterations to its standard molecular weight. Note also that the equivalent mass involved in the molecular bindings leads to non-integer values for the amino acid weights, and that the very weight of each amino acid type is not a single fixed value, but may assume different values, depending on the presence of different isotopes of the various atoms constituting the amino acid. Values reported in Table 1 are just the average masses of the molecules.

| Name | Abbreviations | Molecular Weight | Limitations |
|---|---|---|---|
| Glycine | Gly (or G) | 75.07 | - |
| Alanine | Ala (or A) | 89.34 | - |
| Serine | Ser (or S) | 105.10 | - |
| Proline | Pro (or P) | 115.14 | - |
| Valine | Val (or V) | 117.15 | - |
| Threonine | Thr (or T) | 119.12 | - |
| Cysteine | Cys (or C) | 121.16 | - |
| Taurine | Tau | 125.15 | only c-terminal |
| Piroglutamic Acid | pGlu | 129.10 | only n-terminal |
| Leucine | Leu (or L) | 131.18 | - |
| Asparagine | Asn (or N) | 132.12 | - |
| Aspartic Acid | Asp (or D) | 133.11 | - |
| Glutamine | Gln (or Q) | 146.15 | - |
| Lysine | Lys (or K) | 146.19 | - |
| Glutamic Acid | Glu (or E) | 147.13 | - |
| Methionine | Met (or M) | 149.22 | - |
| Histidine | His (or H) | 155.16 | - |
| Phenylalanine | Phe (or F) | 165.16 | - |
| Arginine | Arg (or R) | 174.21 | - |
| Tyrosine | Tyr (or Y) | 181.19 | - |

Table 1: Commonly considered amino acids.

An accurate and generalizable sequencing procedure should be able to deal with the above uncertainties, by taking as part of the problem data the information about which are the components that should be considered as possible for the current analysis, their weight values, the desired numerical precision of the sequencing procedure, set on the basis of the accuracy of the adopted mass spectrometry device, and any other incidentally known information. When performing an analysis, in fact, we obviously do not know the solution, but we often know which aspects of the solution could be considered as possible for the current analysis, and which ones could not. At worst, if we do not know anything, simply every aspect of the solution should be considered as possible.

This situation may therefore be formalized by considering the number $n$ of possible components (the amino acids) that must be considered for the current analysis, the set $N = \{1, 2, \ldots, n\}$ of the indices $i$ corresponding to such components in increasing weight order, the set

$$A = \{a_1, a_2, \ldots, a_n\}, \qquad a_i \in \mathbb{R}_+$$

of the weight values of such components (the molecular weights of the amino acids) that must be considered for the current analysis, together with the sets

$$\begin{array}{rcl} Min & = & \{m_1, m_2, \ldots, m_n\},\ m_i \in \mathbb{Z}_+ \\ Max & = & \{M_1, M_2, \ldots, M_n\},\ M_i \geq m_i,\ M_i \in \mathbb{Z}_+ \end{array}$$

respectively of the minimum and the maximum of the possible number of molecules of each component that must be considered for the current analysis, the number $d$ of decimal digits that can be considered significant for the current analysis, and a value $\delta \in \mathbb{R}_+$ of the maximum numerical error that may occur in the current analysis.

Amino acids can link to each other into a peptidic chain, by connecting the aminic group $NH_2$ of one molecule with the carboxylic group COOH of another molecule. The free $NH_2$ extremity of the peptide is called N-terminus, while the free COOH extremity is called C-terminus. Some amino acids, expecially the modified ones, can be situated only in particular positions of the sequence, i.e. only N-terminal or only C-terminal. Since each of the peptidic bonds releases an $H_2O$ molecule, the weight of a peptide is not simply the sum of the weights of its component amino acids. Moreover, the weights observed in the spectrum correspond to the actual weights only for the ionized molecules (ions) which retain one single electrical charge. When, on the other hand, a ion retains more than one charge, the weight observed in the spectrum is only a fraction of the actual ion weight. By considering the set

$$Y^0 = \{y^0_1, y^0_2, \ldots, y^0_n\}, \qquad y^0_i \in \mathbb{Z}_+$$

of the numbers of molecules of each component (here the amino acids) contained in the overall polymer (here the peptide), and the number $e_0 \geq 1$ of electrical charges retained by the ionized polymer, the observed weight $w_0$ of the overall polymer is given by the following equation,

$$w_0 = \frac{\sum_{i \in N} \left(y^0_i (a_i - c_a)\right) + c_a + c_0 e_0}{e_0} \pm \delta \tag{1}$$

where $c_a$ and $c_0$ are constant values. When considering $d = 3$ decimal digits, $c_a$ is 18.015 and $c_0$ is 1.008.

**Example 4.1** A small peptide with sequence Leu-His-Cys-Thr-Val ionized by only one charge, considering only $d = 2$ decimal digits, has an observed weight of $w_0 = (131.18 - 18.02) + (155.16 - 18.02) + (121.16 - 18.02) + (119.12 - 18.02) + (117.15 - 18.02) + 19.02 \pm \delta = 572.69 \pm \delta$.

Several different types of fragments can be obtained during the fragmentation process. In particular, there are three possible standard N-terminal ionized fragments, called a-ion, b-ion, c-ion, and three possible standard C-terminal ones, called x-ion, y-ion, z-ion, as illustrated in Fig. 2. Note that b-ions and y-ions are generally the most common.

Again, each fragment has a weight which is not simply the sum of those of its component amino acids. By considering the number $f$ of fragment peaks selected in the spectrum; the set $F = \{1, 2, \ldots, f\}$ of the indices $j$ corresponding to such peaks in decreasing weight order; the set

$$W = \{w_1, w_2, \ldots, w_f\}, \qquad w_j \in \mathbb{R}_+$$

of the weights corresponding to such peaks (so that $w_0$ remains the weight of the overall peptide); the sets

$$Y^j = \{y_1^j, y_2^j, \ldots, y_n^j\}, \qquad y_i^j \in \mathbb{Z}_+ \qquad j = 1, \ldots, f$$

of the numbers of molecules of each component contained in the fragment of weight $w_j$, $j = 1, \ldots, f$; the number $t_{\max}$ of all the possible standard types of fragments that should be considered for the current analysis; the set

$$T = \{1, 2, \ldots, t_{\max}\}$$

of the indices $t$ corresponding to such types; the maximum number of electrical charges $e_{\max}$ that a ion may retain in the current analysis, the set

$$E = \{1, 2, \ldots, e_{\max}\}$$

of the numbers $e$ of electrical charges that a ion may retain in the current analysis; the type $t_j \in T$ of the fragment of weight $w_j, j = 1, \ldots, f$; the number $e_j \in E$ of electrical charges retained by the fragment of weight $w_j$, $j = 1, \ldots, f$, the relation that can be observed in the spectrum is the following.

$$w_j = \frac{\sum_{i \in N} \left[y_i^j(a_i - c_a)\right] + c_t + c_0 e_j}{e_j} \pm \delta, \qquad j = 1, \ldots, f \tag{2}$$

Values $c_a$ and $c_0$ are as above, and $c_t$ is a constant value depending on the type $t_j$ of the fragment. When considering $d = 3$ decimal digits, $c_t$ is -28.002 for a-ions, 0.000 for b-ions, 17.031 for c-ions, 44.009 for x-ions, 18.015 for y-ions, 1.992 for z-ions.

Besides, additional (non standard) fragmentation may also occur: losses of small neutral molecules such as water, ammonia, carbon dioxide, carbon monoxide, or breaking of a side chain. In such cases, the weight of the fragment decreases accordingly. Finally, since fragments appear in the spectrum only when they are ionized, the fact that a fragment is observed does not mean that its complement fragment will be observed as well.

**Example 4.2** When considering the spectrum reported in Fig. 1, and making the simplifying hypothesis of selecting only the peaks labelled with numbers (even if in practice a slightly larger set of peaks should be considered), we have $w_0 = 851.3$, $f = 9$, and $W = \{$ 764.3, 651.3, 627.1, 538.2, 496.1, 425.1, 382.9, 201.0, 173.1 $\}$.

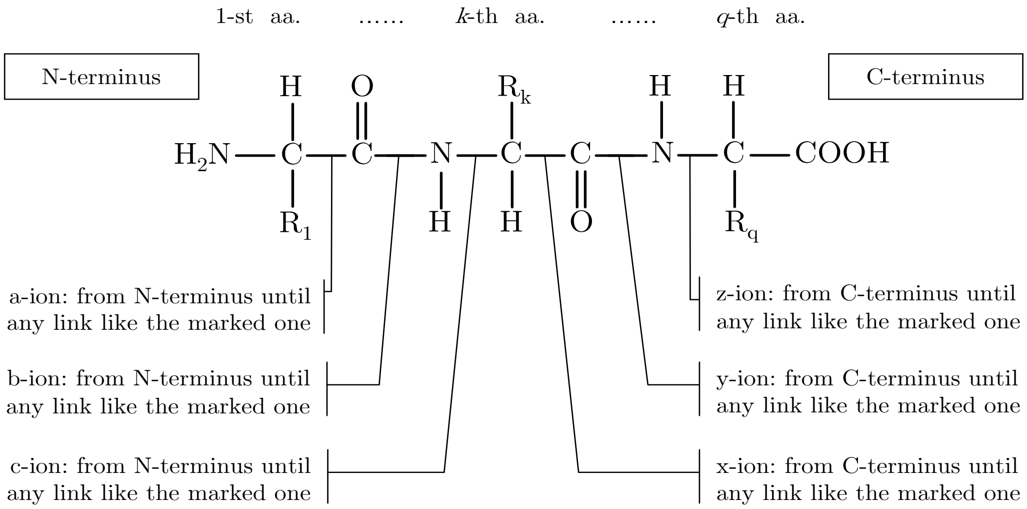

1-st aa.
......
k-th aa.
......
q-th aa.
N-terminus
C-terminus
H
O
R_k
H
H
H_2N
C
C
N
C
C
N
C
COOH
R_1
H
H
O
R_q
a-ion: from N-terminus until any link like the marked one
b-ion: from N-terminus until any link like the marked one
c-ion: from N-terminus until any link like the marked one
z-ion: from C-terminus until any link like the marked one
y-ion: from C-terminus until any link like the marked one
x-ion: from C-terminus until any link like the marked one

Figure 2: Different types of fragments obtainable from a peptide.

## 5 A Logic Encoding of the Peak Interpretation Problem

Each peak of weight $w_j$ selected from the spectrum may be of one of the types $t \in T$, but the exact type is generally unknown. In other words, each peak may have several different *interpretations*. If a peak of weight $w_j$ is considered for instance an a-ion, it may have a certain sequence; if it is considered a b-ion, it cannot have that sequence, and so on. Moreover, since there are rules about incompatibility of fragments and electrical charges of ions, not all of the interpretations are admissible: when interpreting one peak, the interpretations given to all other peaks must be considered. The peak interpretation problem is therefore a decision problem that should be solved by considering all peaks at the same time, and which is defined as follows.

**Definition 5.1** The *peak interpretation problem* consists in assigning to each peak $w_j$ selected from the spectrum, $j = 1, \ldots, f$, (at least) one hypothesis about the type $t_j \in T$ and the charge $e_j \in E$ of the fragment that originated $w_j$ in such a way that all interpretations given to all peaks are coherent. Coherent means that they respect a number of *rules* formalizing our knowledge of the problem.

Rules holding for every analysis are the incompatibility and multicharge rules given below. Other analysis-specific rules may be generated, as observed below. Note that each peak should have *at least* one interpretation, but not necessarily *only* one. A peak may in fact be originated by more than one type of fragment incidentally having the same observed weight, even if this happens very rarely in practice.

We formalize the peak interpretation problem by means of propositional logic. By denoting with $w_j \rightarrow t, e$ the fact that peak $w_j$ is interpreted as being due to a fragment of type $t \in T$ and having an electrical charge $e \in E$, we consider for each interpretation of $w_j$ a propositional variable

$$x_{j \rightarrow t,e} \in \{\mathit{True}, \mathit{False}\}, \qquad j \in F,\ t \in T,\ e \in E$$

When considering for instance the 6 above standard types of fragments obtainable from a peptide and a maximum electrical charge $e_{\max} = 2$, we have $T = \{1, 2, 3, 4, 5, 6\}$ and $E = \{1, 2\}$. The possible interpretations of a peak $w_j$ are therefore 12, and this may be represented by means of the following clause containing 12 variables

$$(x_{j\rightarrow 1,1} \vee x_{j\rightarrow 2,1} \vee \ldots \vee x_{j\rightarrow 6,1} \vee x_{j\rightarrow 1,2} \vee x_{j\rightarrow 2,2} \vee \ldots \vee x_{j\rightarrow 6,2})$$

In order to get rid of the fact that the weight of peptides and of their fragments is not simply the sum of those of their component amino acids, we define now a different (theoretical) model of polymeric compound, as follows.

**Definition 5.2** Given a (real) single charge peptide of observed weight $w_0$, the *normalized peptide* associated with it is a (theoretical) polymeric compound of weight $w_0 - (c_a + c_0)$. The possible components of such normalized peptide are (theoretical) components having the following weights (which are those that amino acids assume in the internal part of the peptidic chain)

$$\bar{A} = \{(a_1 - c_a), (a_2 - c_a), \ldots, (a_n - c_a)\}$$

As a result, the weight of the normalized peptide, as well as the weights of its fragments, is simply the sum of those of its components. By the above definition, the normalization of a single charge real peptide of observed weight $w_0$ is composed by a number of molecules of each of the components in $\bar{A}$ equal to the number of molecules $Y^0 = \{y^0_1, y^0_2, \ldots, y^0_n\}$ of each amino acid contained in the real peptide of observed weight $w_0$.

**Example 5.3** The normalized peptide corresponding to the real peptide of weight 572.69 of Example 4.1 has a weight of (572.69 - 19.02) = 553.67, and its component have the following weights: (131.18 -18.02) = 113.16, (155.16 -18.02) = 137.14, (121.16 -18.02) = 103.14, (119.12 -18.02) = 101.10, (117.15 -18.02) = 99.13. If such normalized peptide breaks for instance in Leu-His and Cys-Thr-Val, such fragments respectively have weights: (113.16 + 137.14) = 250.30 and (103.14 + 101.10 + 99.13) = 303.37.

We will consider for such normalized peptide the above described topological concepts of N-terminus, C-terminus, peptidic bonds, etc., in their intuitive sense, as if it was a real peptide.

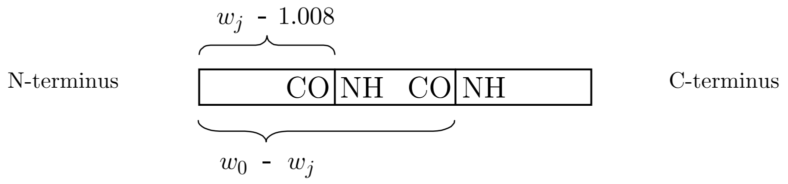
When a peak receives an interpretation, this means that an hypothesis has been done about where the cleavage occurred in the peptide, and also about which was the chemical structure of the peptide in that point. Asserting that, for a single charge peptide of observed weight $w_0$, peak $w_j$ is, for instance, a single charge b-ion means that, starting from the N-terminus of the normalization of that peptide, there has been a cleavage between CO and NH, and that the part of such normalization going from the N-terminus to that cleavage has a weight of

$$w_j - 1.008 \pm \delta$$

On the contrary, asserting that, for the same single charge peptide of observed weight $w_0$, the same peak $w_j$ is now, for instance, a single charge y-ion means that, starting from the C-terminus of the normalization of that peptide, there have been a cleavage between NH and CO, and that the part of such normalization going from the C-terminus to that cleavage has a weight of $w_j - 19.023 \pm \delta$. Therefore, the part of the same normalization going from the N-terminus to that cleavage has a weight of

$$w_0 - (c_a + c_0) - (w_j - 19.023) \pm \delta = w_0 - w_j \pm \delta$$

The two interpretations therefore bring to radically different hypothesis on the structure of the normalized peptide, as illustrated by the following diagram for $w_0 - (c_a + c_0) \approx 850$ and $w_j \approx 300$.



We now consider, for the each variable $x_{j \to t,e}$, with $j \in F,\ t \in T,\ e \in E$, the weight that the part of the normalized peptide going from the N-terminus to the cleavage corresponding to interpretation $w_j \to t, e$ would assume.

**Definition 5.4** An *N-terminal portion* of a normalized peptide is any part of that compound going from the N-terminus to any peptidic bond between CO and NH (a part that, if such bond was broken, would constitute a b-ion). The *hypothesized weight* of such N-terminal portion is the one given by the following function $b(j, t, e)$

$$b(j,t,e) = \begin{cases} (w_j - c_t - c_0 e_j) e_j & \text{for a-ions, b-ions, c-ions} \\ (w_0 - c_a - c_0 e_0) e_0 - (w_j - c_t - c_0 e_j) e_j & \text{for x-ions, y-ions, z-ions} \end{cases}$$

Note that charge $e_0$ of the precursor ion is known and fixed during each single analysis. By using the above concepts, variable $x_{j \to t,e} = True$ implies that there exists an N-terminal part of the normalized peptide having weight $b(j,t,e) \pm \delta$.

$$x_{j \to t,e} = True \quad \Rightarrow \quad \text{N-terminus} \quad \underbrace{\boxed{\qquad \text{CO}}}_{b(j,t,e)}\!\boxed{\text{NH} \qquad} \quad \text{C-terminus}$$

We are now able to introduce, in form of clauses, the additional sets of rules that an interpretation should respect in order to be coherent. A first one is the set of *incompatibility* rules. To this aim, we denote here variables using their corresponding values for $b$. Two variables $x_{b'}$ and $x_{b''}$ are incompatible if, for example, the difference between $b'$ and $b''$ is smaller than the smallest possible component, that is:

$$|b' - b''| < (a_1 - c_a) - 2\delta$$

More generally, $x_{b'}$ and $x_{b''}$ are incompatible if the difference between $b'$ and $b''$ has a weight value which cannot be any combination of possible components. In other words, it does not exist any non-negative integer vector $(y_1, y_2, \ldots, y_n)^{tr} \in \mathbb{Z}_+^n$ verifying the following equation.

$$|b' - b''| = y_1(a_1 - c_a) + y_2(a_2 - c_a) + \ldots + y_n(a_n - c_a) \pm 2\delta$$

Therefore, incompatibility clauses of the following form are added for all the couples of incompatible variables $x_{b'}$ and $x_{b''}$.

$$(\neg x_{b'} \vee \neg x_{b''})$$

Another set of rules that should be considered in order to have a coherent interpretation is that of *multicharge* rules. Depending on the mass spectrometry device, ions retaining more than one electrical charge, called multicharged ions, are usually less common than single charged ions, and it is common practice to assume that, if a multicharged ion has been observed in the spectrum, also the corresponding single charged one should appear in the spectrum. Therefore, each variable $x_{j' \to t,e}$ with $e > 1$ implies, if it exists, another variable $x_{j'' \to t,1}$ with $(j' - c_0 e)e = j'' - c_0$, as follows

$$(\neg x_{j' \to t,e} \vee x_{j'' \to t,1})$$

Finally, a number of additional clauses representing a priori known information about the specific mass spectrometry device used for the analysis, about the analyzed compound, or about other possibly known relations among the interpretations of the various peaks may also be generated. This because, clearly, the more information can be introduced by means of clauses, the more reliable the results of the analysis will be.

By assuming no limitations on the structure of the generated clauses, therefore allowing the full expressive power of propositional logic, we obtain at this point a set of $v$ clauses $C_1, C_2, \ldots, C_v$. Generally, incompatibility clauses are by far the more numerous. Since all clauses must be considered together, we construct their conjunction, that is a generic propositional formula $\mathcal{F}$ in *conjunctive normal form* (CNF)

$$\mathcal{F} = C_1 \wedge C_2 \wedge \ldots \wedge C_v$$

Each truth assignment $\{$*True,False*$\}$ for the variables $x_{j \rightarrow t,e}$, with $j \in F$, $t \in T$, $e \in E$, such that $\mathcal{F}$ evaluates to *True* is known as a *model* of $\mathcal{F}$. We now have the following result.

**Theorem 5.5** Each model $\mu$ of the generated propositional formula $\mathcal{F}$ is a coherent solution of the peak interpretation problem for the peptide under analysis. Moreover, no coherent solution of the peak interpretation problem which does not correspond to a model $\mu$ of $\mathcal{F}$ can exist.

The proof relies in the fact that the formula $\mathcal{F}$ represents by construction all the rules (peak assignment rules, incompatibility rules, multicharge rules) that a peaks interpretation must satisfy to be considered coherent. Therefore, each model $\mu$ is an interpretation satisfying all the rules. Conversely, each interpretation satisfying all the rules corresponds to a truth assignment for the variables $x_{j \rightarrow t,e}$ such that $\mathcal{F}$ is *True.*

Finding a model of a generic CNF, or proving that such model does not exist, is known as the *satisfiability* problem (SAT). Extensive references can be found in [9, 15, 17, 28]. This problem is NP-complete [14] in the general case. However, for the average size of generated instances, solution times of a DPLL branching algorithm are very moderate. Note also that, in some special cases of peptide analysis, one may be able to obtain polynomially solvable formulae by imposing syntactical limitations on the structure of the generated clauses (see e.g. [3, 8, 12, 20]). For instance, when considering only b-ion and y-ion as the possible types of fragments, and only single charged ions, we obtain Quadratic formulae [1], which are polynomially solvable.

Since we are interested in all possible solutions of the peptide analysis, we are interested in all the possible peaks interpretations, that is we are interested in finding all the models

$$\{\mu_1, \mu_2, \ldots, \mu_r\}$$

of $\mathcal{F}$. This was obtained in practice by modifying the SAT solver BrChaff [6] in such a way that, after finding a model, the search does not stop, but keeps exploring the branching tree, until its complete examination.

In the case $\mathcal{F}$ does not even have one model, this may mean that the considered sets of fragment types $T$ and/or possible charges $E$ are not enough to give an interpretation to every considered peak, or simply that the mass spectrometry analysis suffered from some experimental disturbance which produced uninterpretable noise peaks. In such latter case, either the mass spectrometry should be improved, or the formula $\mathcal{F}$ should be considered as an instance of the *maximum satisfiability* problem (Max-SAT) [28], which consists in finding a truth assignment for the variables $x_{j\to t,e}$ maximizing the number of clauses which evaluate to *True*. Note that this latter solution means that not all rules for having a coherent interpretation are respected, therefore the result of the analysis is less reliable.

**Example 5.6** When considering the compound of Example 4.2. ($w_0 = 851.3$, $f = 9$, and $W = \{764.3, 651.3, 627.1, 538.2, 496.1, 425.1, 382.9, 201.0, 173.1\}$), the possible components of Table 1, and allowing a-ion, b-ion, c-ion, x-ion, y-ion, z-ion, and double and single charges, we obtain a formula $\mathcal{F}$ with 108 variables and 4909 clauses, which has 3 models.

It is worth to note that the SAT problem, and all its variants above described, can be solved not only working in the field of propositional logic (as it is done by BrChaff and many other solvers), but also working with Integer Linear Programming. Each clause, written in the following general form ($P$ is the set of indices of positive variables, $N$ that one of negative variables)

$$\bigvee_{k\in P} x_k \vee \bigvee_{k\in N} \neg x_k$$

can be converted into the following linear inequality

$$\sum_{k\in P} x_k + \sum_{k\in N} (1 - x_k) \geq 1$$

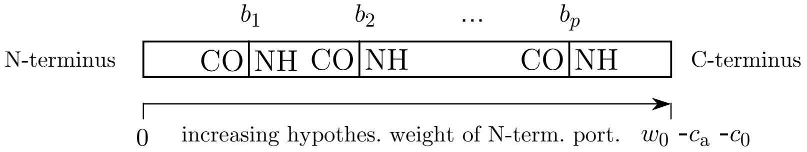
Therefore, the set of all clauses becomes a set of linear inequalities constituting the constraints of the ILP, an objective function can be added, and algorithms for solving ILP can now be used. Generally speaking, however, the complexity of solving the above described problems does not change: when the SAT problem belongs to an easy special class, the same happens for the ILP. See e.g. [9] for further details.

# 6 Computing the Weights Database and Generating the Sequences

As described, each variable $x_{j\to t,e}$ with $j \in F$, $t \in T$, $e \in E$, corresponds to an hypothesized weight $b(j, t, e)$ of an N-terminal portion of the normalized peptide. Therefore, given a model $\mu$ for the generated formula $\mathcal{F}$, consider all the hypothesized weights of the N-terminal portions corresponding to all the *True* variables of $\mu$. By ordering such values in increasing weight order, we obtain what we call the *succession of breakpoints* $B^\mu$ corresponding to model $\mu$ for the normalized peptide under analysis.

$$B^\mu = \{b_1, b_2, \dots, b_p\}$$

This means that, when giving to the considered peaks $W$ the interpretation represented by $\mu$, we have located the peptidic bonds of the normalized peptide under analysis at the locations given by the values of the elements of $B^{\mu}$, as illustrated by the following diagram.


$b_1$
$b_2$
$\dots$
$b_p$
N-terminus
CO
NH
CO
NH
CO
NH
C-terminus
0
increasing hypothes. weight of N-term. port.
$w_0$ -$c_{\mathrm{a}}$ -$c_0$

**Definition 6.1** Define now a *gap* as the difference between two adjacent breakpoints $(b_{h+1}, b_h)$, and a corresponding *subsequence* as the portion of the normalized peptide spanning between the two peptidic bonds corresponding to the two above adjacent breakpoints $(b_{h+1}, b_h)$.

Now we compute, for each value of gap $b_{h+1} - b_h$, all the non-negative integer vectors $(y_1, y_2, \dots, y_n)^{tr} \in \mathbb{Z}^n_+$ verifying the following equation.

$$b_{h+1} - b_h = y_1(a_1 - c_a) + y_2(a_2 - c_a) + \dots + y_n(a_n - c_a) \pm 2\delta$$

The results are all the possible subsequences that may cover the gap $b_{h+1} - b_h$. Denote such set of subsequences by $S(b_{h+1} - b_h)$. Note that $S(b_{h+1} - b_h)$ depends only on the value of the gap $b_{h+1} - b_h$, not on the locations of the breakpoints. The first gap $b_1 - 0$ and the last one $w_0 - (c_a + c_0) - b_p$ should be managed in a way which is slightly different from that of the central gaps. They are indeed the only gaps which may contain components having limitation on their position in the sequence (only N-terminal or only C-terminal, see Section 2), hence this should be considered. Furthermore, only an imprecision $\delta$ instead of $2\delta$ should be considered for the first gap, since only one extremity of the gap can be affected by such imprecision. Define $b_0 = 0$ for a more uniform notation.

In order to compute such subsequences, we use a *weights database*, as follows. The possible components of the normalized peptide can be view as an alphabet $\Sigma$ on $n$ symbols. For instance, if the possible components are the 20 amino acids reported in Table 1, we have

$$\Sigma = \{\mathrm{Gly}, \mathrm{Ala}, \dots, \mathrm{Tyr}\}$$

A subsequence of the normalized peptide is just a sequence of components, and therefore a string over this alphabet. Its weight is normalized, and therefore can be computed by summing the weights of the components. The set of all such strings may be denoted as $\Sigma^*$. Knowing the correspondences between all the elements of $\Sigma^*$ and their weights would of course speed-up the operation of finding the subsequences. However, generating all $\Sigma^*$ would be clearly impossible from a computational point of view. On the other hand, the set of strings having a molecular weight not greater than $\lambda$ may be denoted as $\Sigma^{*\leq\lambda}$. If $\lambda$ is greater than or equal to the maximum of the mentioned gaps, also $\Sigma^{*\leq\lambda}$ may give the same help in the operation of finding the subsequences. For useful values of $\lambda$, however, $\Sigma^{*\leq\lambda}$ generally becomes too large.

We propose a procedure to consider it implicitly. Not that $\Sigma^{*\leq\lambda}$, for any fixed $\lambda$, can be computed by using only the information about the possible components for the current analysis (or better yet, for the set of current analyses). We therefore compute it off-line, before starting any sequence analysis, as soon as the information about the possible components is available. Every sequence is put in correspondence with a natural number, by considering the components of the sequence as a number expressed in base $n+1$ ($n$ is the number of components). This correspondence must be biunivocal and easily computable. For instance, with the 20 amino acids reported in Table 1, considering the sequence written horizontally, the last (the rightmost) element would correspond to the symbol multiplying $21^0$, the last-but-one element would correspond to the symbol multiplying $21^1$, and so on. Moreover, the first symbol (Gly) in the list of possible components (Table 1) would mean number 1, the second (Ala) number 2, and so on. An empty position (no amino acid) would mean number 0. This because, if any other amino acid would mean 0, a sequence beginning with that amino acid would correspond to the same number as the same sequence without the initial amino acid, and the correspondence would not be biunivocal.

**Example 6.2** The sequence Gly-Ser-Gly-Tyr, or, more precisely,

< no amino acid> ... < no amino acid > Gly Ser Gly Tyr

would then corresponds to the number 0 ... 0 1 3 1 20(or K) in base 21, that in base 10 is $20 \times 21^0 (= 20) + 1 \times 21^1 (= 21) + 3 \times 21^2 (= 1323) + 1 \times 21^3 (= 9261) = 10625$.

The weights of all sequences up to molecular weight $\lambda$, are therefore computed off-line, and stored in correspondence with the described natural numbers representing the sequences. This computation may be done efficiently using smaller solutions to gradually compute larger solutions. Note that more sequences may have the same molecular weight, so one weight may correspond to more than one natural number, even if one natural number corresponds to only one sequence, hence to one weight. The natural numbers may also be not stored, but simply be the indices of an array memorizing the weights. This constitutes the weights database: given a molecular weight, it allows to find almost instantaneously which are all the sequences of components that could produce a portion of normalized peptide having that weight. Value $\lambda$ is chosen big enough to cover all the possible gaps that one could need to sequence in the set of current analyses.

Therefore, for each gap $b_{h+1} - b_h$, the set of all the possible subsequences $S(b_{h+1} - b_h)$ covering that gap is computed in extremely short times by searching the weights database for all natural numbers corresponding to the weight $b_{h+1} - b_h$, and by explicitly generating the subsequences corresponding to such natural numbers.

When all the sets of subsequences $S(b_{h+1} - b_h),\ h = 0, \ldots, p$ are available, all the possible sequences $\mathcal{S}_\mu$ of the normalized peptide under the peak interpretation $\mu$ can be generated with the concatenation of such sets in all possible ways, operation which we denote by $\oplus$, but eliminating sequences violating the requirements regarding minimum $m_i$ or maximum $M_i$ value on the number of each component.

$$\mathcal{S}_\mu = S(b_1 - b_0) \oplus S(b_2 - b_1) \oplus \ldots \oplus S(w_0 - c_0 - b_p)$$

Finally, when considering the sets of all the possible sequences $\{\mathcal{S}_{\mu_1}, \mathcal{S}_{\mu_2}, \ldots, \mathcal{S}_{\mu_r}\}$ for all the possible models $\{\mu_1, \mu_2, \ldots, \mu_r\}$ of $\mathcal{F}$, the complete set of all possible sequences $\mathcal{S}$ of the normalized peptide is obtained:

$$\mathcal{S} = \mathcal{S}_{\mu_1} \cup \mathcal{S}_{\mu_2} \cup \ldots \cup \mathcal{S}_{\mu_r}$$

By construction, the set of all the possible sequences $\mathcal{S}$ of the normalized peptide is also the set of all the possible sequences of the real peptide under analysis, so the sequencing problem have been solved.

Note that, in the case when the formula $\mathcal{F}$ is unsatisfiable, and a truth assignment maximizing the number of clauses which evaluates to *True* has been found, some gap may admit no subsequences because some incompatibility clauses are not respected. A less reliable solution can in this case be obtained by merging each unsequenceable gap with one of its neighbouring ones (preferably the smaller).

**Example 6.3** When considering the formula $\mathcal{F}$ of Example 5.6 with 108 variables, 4909 clauses and 3 models, computing the weights database with $\lambda = 300$ we obtain 3 breakpoint successions, reported below together with all their corresponding possible sequences:

{87.0, 224.2, 339.2, 452.2, 565.2, 662.2} which gives two sequences:
Ser-His-Asp-Leu-Leu-Pro-Gly-Leu
Ser-His-Asp-Leu-Leu-Pro-Leu-Gly

{87.0, 224.2, 339.2, 452.2, 565.2, 678.3} which gives two sequences:
Ser-His-Asp-Leu-Leu-Leu-Gly-Pro
Ser-His-Asp-Leu-Leu-Leu-Pro-Gly

{87.0, 184.0, 355.2, 452.2, 565.2, 662.2} which gives four sequences:
Ser-Pro-Gly-Asn-Pro-Leu-Pro-Gly-Leu
Ser-Pro-Gly-Asn-Pro-Leu-Pro-Leu-Gly
Ser-Pro-Asn-Gly-Pro-Leu-Pro-Gly-Leu
Ser-Pro-Asn-Gly-Pro-Leu-Pro-Leu-Gly

However, since in this series of examples we selected from the spectrum of Fig. 1 only the labelled peaks, results are not as accurate as it would be possible when selecting more peaks.

# 7 Implementation and Computational Experience

The proposed approach is implemented in C++ and tested on a Pentium IV 1.7GHz PC. The initial input routine (i) reads all informations about possible components and possible types of fragments and charges and computes the weights database, (ii) reads the spectrum and extracts from it all peaks above a certain value. After this, the logic formula $\mathcal{F}$ representing the peak interpretation problem is generated. All models of $\mathcal{F}$ are then found by means of the DPLL SAT solver BrChaff [6], modified in order to search for all the models of the given formula. Then, for each model $\mu$ of $\mathcal{F}$, the breakpoint succession is computed, and all the possible subsequences covering each gap are computed and linked together.

| Input Data | | | | | Outcomes | | | | Times | |
|---|---|---|---|---|---|---|---|---|---|---|
| $w_0$ | $f$ | $t_{\max}$ | $e_{\max}$ | $n$ | $x$ | $v$ | $r$ | $\mathcal{S}$ | w/o WD | w WD |
| 572.20 | 7 | 2 | 1 | 20 | 14 | 108 | 1 | 1 | 0.1 | 0.1 |
| 572.20 | 7 | 6 | 2 | 20 | 84 | 3571 | 2 | 2 | 1.9 | 1.6 |
| 851.30 | 18 | 2 | 1 | 20 | 36 | 543 | 1 | 4 | 0.5 | 0.5 |
| 851.30 | 18 | 4 | 2 | 24 | 144 | 6780 | 4 | 7 | 2.0 | 1.4 |
| 851.30 | 18 | 6 | 3 | 24 | 324 | 12642 | 10 | 16 | 5.6 | 3.0 |
| 859.12 | 20 | 3 | 1 | 40 | 60 | 2904 | 4 | 26 | 1.6 | 1.1 |
| 859.12 | 20 | 6 | 2 | 40 | 240 | 8156 | 5 | 29 | 4.1 | 3.4 |
| 913.30 | 16 | 2 | 1 | 20 | 32 | 539 | 2 | 7 | 1.0 | 0.8 |
| 913.30 | 16 | 6 | 3 | 20 | 288 | 10741 | 8 | 32 | 6.8 | 4.0 |
| 968.58 | 19 | 2 | 1 | 20 | 38 | 768 | 6 | 24 | 1.3 | 1.1 |
| 968.58 | 19 | 6 | 2 | 20 | 228 | 7021 | 10 | 38 | 4.1 | 3.4 |
| 1037.10 | 18 | 2 | 1 | 20 | 36 | 714 | 7 | 25 | 1.4 | 1.0 |
| 1037.10 | 18 | 6 | 2 | 20 | 216 | 6936 | 12 | 44 | 4.3 | 3.2 |
| 1108.60 | 21 | 2 | 1 | 26 | 42 | 2687 | 8 | 18 | 3.5 | 2.1 |
| 1108.60 | 21 | 4 | 2 | 26 | 168 | 7456 | 16 | 64 | 12.2 | 5.6 |
| 1234.20 | 19 | 2 | 2 | 20 | 76 | 4529 | 9 | 26 | 8.3 | 3.2 |
| 1234.20 | 19 | 6 | 2 | 20 | 228 | 8956 | 15 | 106 | 29.2 | 14.0 |
| 1479.84 | 20 | 2 | 1 | 20 | 40 | 690 | 7 | 22 | 14.3 | 6.8 |
| 1479.84 | 20 | 6 | 2 | 20 | 240 | 8796 | 18 | 102 | 33.9 | 13.7 |
| 1570.60 | 22 | 2 | 1 | 21 | 44 | 2498 | 9 | 35 | 28.5 | 16.3 |
| 1570.60 | 22 | 6 | 2 | 21 | 264 | 9657 | 14 | 98 | 56.8 | 39.2 |
| 1607.69 | 27 | 2 | 2 | 26 | 108 | 5744 | 6 | 20 | 44.3 | 20.9 |
| 1607.69 | 27 | 6 | 3 | 26 | 486 | 22565 | 11 | 63 | 473.0 | 192.8 |

Table 2: Real-world peptide sequencing problems.

Those subsequences may be produced either by means of a specialized branching algorithm working on-line, or by means of the weights database computed offline and used on-line. Finally, by considering the union of the set of sequences corresponding to the different models of $\mathcal{F}$, all the solutions of the sequencing problem are obtained.

Table 2 reports various experiments of real peptide sequencing problems. In particular, we indicate: the weight of the peptide ($w_0$); the number of peaks extracted from the spectrum ($f$); the number of considered types ($t_{\max}$) and charges ($e_{\max}$) of fragments; the number of possible components ($n$); the number of variables ($x$) and clauses ($v$) of the obtained formula; the number of models ($r$) of the obtained formula, the overall number of solutions ($\mathcal{S}$), and computational times (in seconds) for the whole sequencing procedure without the weights database (w/o WD) and with it (w WD). Time for computing off-line the weights database with $\lambda = 300$ is 40 seconds, with $\lambda = 400$ is 126 seconds. Both values were sufficient for sequencing the gaps in the reported analyses. A time of this order (the exact one depends on our *a priori* choice for $\lambda$) should therefore be considered just once for a whole series of tests with WD. It can also be stored on hard disk and read by the input routine in a subsequent time. Those results are intended to give real-world examples of application, rather than exploring all the computational possibilities of the proposed procedure.

As observable from the table, results depend of course on the choice of possible types and charges of fragments: for the same spectrum, different choices produce different results, and the number of sequences compatible with the given input data is sometimes large. This is an intrinsical character of the problem. However, all the solutions are generally very related, in the sense that some parts are just common, and some other are given by all the combinations of a (generally small) number of components.

The use of the weights database is always able to reduce computation times. This reduction increases when increasing the solution time, and grows faster than the latter one. In the examples, it passes from about 0.2 sec. for a problem with solution time of 1 sec., i.e. a reduction of 20%, to about 280 sec. for a problem with solution time of 473 sec., i.e. a reduction of 59%. Therefore, the more consistent speed-ups are obtained for the larger instances (the ones for which they are more useful). The whole procedure, according to biochemist experts, is a very powerful, accurate and flexible sequencing tool, and allows the sequencing of compounds not handled by other available techniques.

# 8 Conclusions

The problem of the determination of the amino acid sequence of a peptide is considered. Such problem is of basic relevance in biological and medical research, but is difficult to model and computationally hard to solve. Data obtained from the mass spectrometry analysis of a generic polymeric compound, constituted, according to specific chemical rules, by a sequence of components, are here used to build a propositional logic formula. The models of this formula represent coherent interpretations of the set of data, and are employed to generate all possible correct results of the analysis itself. The problem has been therefore subdivided into a *peaks interpretation* phase and a *sequence generation* phase.

The peaks interpretation phase is solved by means of a DPLL SAT solver modified in order to search for all the models of a formula. The sequence generation phase is solved by computing off-line a weights database, so that all sequences up to a certain molecular weight can be considered implicitly, but only the needed ones generated explicitely. Results of tests on real-world peptide sequencing problems demonstrate the effectiveness of this approach. The use of the weights database is able to sensibly reduce computation times, especially for larger instances.

**Acknowledgements** 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, to Prof. L. De Angelis and Mr. F. Giavarini for the mass spectrometry experimental analysis, and to dr. Antonello Moscatelli for his contribution to the implementation work.

# References

[1] B. Aspvall, M.F. Plass, and R.E. Tarjan. A linear time algorithm for testing the truth of certain quantified boolean formulas. *Information Processing Letters* 8, 121-123 (1979).

[2] V. Bafna 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).

[3] E. Boros, Y. Crama, and P.L. Hammer. Polynomial time inference of all valid implications for Horn and related formulae. *Annals of Mathematics and Artificial Intelligence* 1, 21-32 (1990).

[4] R. Bruni. Solving Peptide Sequencing as Satisfiability. *Computer and Mathematics with Applications* 55(5), 912-923 (2008).

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

[6] R. Bruni and A. Santori. Adding a New Conflict-Based Branching Heuristic in two Evolved DPLL SAT Solvers. In *Proceedings of the Seventh International Conference on Theory and Applications of Satisfiability Testing* SAT2004 (2004).

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

[8] V. Chandru and J.N. Hooker. Extend Horn clauses in propositional logic. *Journal of the ACM* 38, 203-221 (1991).

[9] V. Chandru and J.N. Hooker. *Optimization Methods for Logical Inference.* Wiley, New York (1999).

[10] T. Chen, 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).

[11] W.F. Clocksin. Logic programming and digital circuit analysis. *Journal of Logic Programming*, 4(1), 59-82 (1987).

[12] M. Conforti and G. Cornuéjols. A class of logical inference problems soluble by linear programming. *Journal of the ACM* 42(5), 1107-1113 (1995).

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

[14] M.R. Garey and D.S. Johnson. *Computers and Intractability.* Freeman, New York (1979).

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

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

[17] H. Kleine Büning and T. Lettman. *Propositional logic: deduction and algorithms.* Cambridge University Press, Cambridge (1999).

[18] T.D. Lee. Fast atom bombardment and secondary ion mass spectrometry of peptides and proteins. In *Methods of Protein Microcharacterization*, J.E. Shively (editor) 403-441, Humana Press, Clifton, NJ (1986).

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

[20] J.S. Schlipf, F.S. Annexstein, J.V. Franco, and R.P. Swaminathan. On Finding Solutions for Extended Horn Formulas. *Information Processing Letters* 54(3), 133-137 (1995).

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

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

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

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

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

[26] J.T. Stults. Peptide sequencing by mass spectrometry. *Method Biochem. Anal.* 34, 145-201 (1990).

[27] J.A. Taylor and R.S. Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* 73, 2594-2604 (2001).

[28] K. Truemper. *Effective Logic Computation.* Wiley, New York (1998).

[29] P. Van Hentenryck. *Constraint satisfaction in logic programming.* MIT Press, Cambridge, MA (1989).