Solving Peptide Sequencing as Satisfiability [1]

Renato Bruni

Dip. di Ingegneria Elettronica e dell'Informazione, Università di Perugia, Via G. Duranti, 93 - I-06125 Perugia - ITALY

Abstract

This paper presents an approach for determining the amino acid sequence of a peptide through the solution of propositional satisfiability problems. Data obtained from the mass spectrometry analysis of a peptide are used to build a propositional logic formula, whose models represent coherent interpretations of that set of data and can be used to generate all possible correct results of the analysis itself. Some computational results on real-world peptide analysis problems are reported, which show the effectiveness of our approach.

Key words: De Novo Sequencing, Mass Spectrometry, Propositional Logic Modelling.

1 Introduction

Proteins are composed by the joining of smaller parts called *peptides*, while each peptide is composed by a single sequence of components called *amino acids*. The analysis of the amino acid sequence of peptides, called *sequencing*, is one of the most important and frequent issues in biological and medical research. In particular, protein analyses are generally achieved by dividing a protein molecule into its component peptides (via enzymatic digestion and subsequent fractionation with HPLC or capillary electrophoresis), and by individually analyzing each peptide. Thus, peptide sequencing arises as a fundamental step in protein identification. Moreover, peptide sequencing has an importance on its own in a number of situations such as the study of unknown

Email address: bruni@diei.unipg.it (Renato Bruni).

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

peptides, the research for new drugs, and the synthesis of peptide-like active factors and peptides used in therapy.

A first approach to peptide sequencing was the so-called Edman method [1], which may be implemented either manually or through the use of automatic devices called protein sequenators. However, such a procedure has several drawbacks [2]. Nowadays, a widely used and well established approach to peptide sequencing consists in the use of mass spectrometry (e.g. [3–5]). Such kind of analysis produces a *mass spectrum*, that is the absolute molecular weight distribution of the molecules of a sample containing the peptide under analysis. The study of the weight pattern in the spectrum can be used for understanding the peptide sequence (e.g. [6]). The sequencing is generally further helped by the use of the so called MS/MS (mass spectrometry/mass spectrometry), or tandem mass, methodology (e.g. [7]). 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 a new mass analysis. By doing so, the analysis gives the absolute molecular weight of the full molecules of the precursor ion, as well as those of the various ionized fragments that could be obtained from such kind of molecules. Note that, on the contrary, non ionized molecules do not appear in the spectrum. Such experiments are performed by using several instrumental configurations, mainly triple quadrupole (QQQ), quadrupole time-of-flight (Q-TOF) and ion trap devices [5]. Since the weights of the possible amino acid components are known, and rules for determining the weights of amino acid sequences of known composition are available (even if, unfortunately, the weight of a sequence is not simply the sum of the weights of the components), one could in principle use the MS/MS information in order to determine the sequence. Note, also, that the molecular weights of the above-mentioned fragments of the peptide constitute an essential information for the sequencing.

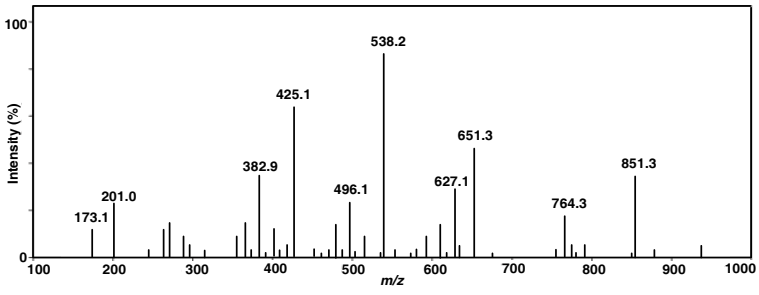


Fig. 1. A MS/MS spectrum generated by collision-induced dissociation.

A typical MS/MS spectrum, however, does not contain any direct reference to amino acids, being a mere succession of peaks corresponding to different molecular weights (see e.g. Fig. 1). Further processing is then requested, and

generally performed as follows. To begin with, all peaks below a certain intensity are removed, being too noise-affected to be considered significant. After this *peak selection* phase, the higher molecular weight is assumed to be the one of the full peptidic complex under analysis, whereas the others correspond to its fragments. Now, a known approach consists in looking for peptide-specific weight patterns in the spectrum (called peptide tags, or fragment fingerprints), and checking them against similar patterns available from data bases [8]. The use of data bases assumes that the protein (or the peptide) under investigation is an already known one. However, due to the very large number of possible sequences, this is not a frequent case. Moreover, a protein may also differ from the standard known form because the sequence underwent some modifications. Therefore, alternative methods are often required and direct identification is to be addressed.

Direct peptide sequencing, known as *de novo* sequencing, is achieved by various recently available techniques (many of which developed by mass spectrometry producers). These procedures: (i) either look for continuous series of fragments differing by just one amino acid, which is therefore identified, or (ii) iteratively generate a large number of virtual sequences and evaluate the match of the corresponding (theoretical) mass patterns with the (actual) mass pattern of the peptide under investigation. In both cases, the whole sequence can be obtained when the spectrum contains the complete series of the fragments. This, however, is often unlikely to occur. The fragmentation process is a stochastic one, and though in fact peptides tend to break at the conjunction of amino acids, they usually do not break at every conjunction of amino acids. Furthermore, such cleavages may be of several different types. And, if the intensity of the hitting is increased, the peptide produces more fragments, but may break at locations which are not the conjunction of amino acids. This makes the problem a very difficult one for the above de novo techniques. Note, moreover, that there are also cases when the information contained in the spectrum is simply not enough to determine a unique sequence, because more than a sequence exists which perfectly fits such a spectrum. Consider, for instance, the case of an incomplete fragmentation: it would be impossible to determine the exact sequence of a peptide portion which did not break up. In these cases, all the possible sequences fitting the spectrum should be found, in order to guarantee accurate and objective results of the analysis.

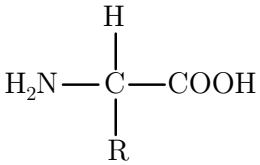
Due to its combinatorial nature, the problem has been more recently approached by means of two different discrete mathematics approaches. The first is the graph theoretical construction proposed in [9], which evolved into the dynamic programming algorithms proposed in [10,11]; the second is the branching-based algorithm proposed in [12]. The first approach has the advantage of requiring polynomial computational time for finding each solution [13], when imposing some limitations to the problem, namely no multicharged fragments can appear in the spectrum, and only peaks corresponding to a set

of fragment types which is "simple" [11] (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 second approach, on the contrary, has no structural limitations regarding types of fragments and charges, and performs a complete search. However, it produces an heavier computational load, and so it is hardly applicable for peptide above a certain dimension.

This paper presents an innovative construction for approaching the peptide sequencing problem as a propositional satisfiability problem. Data obtained from the mass spectrometry analysis of a generic peptide, constituted by an unknown sequence of amino acids, can be used to build a propositional logic formula. The models of this formula can be employed to generate all possible correct results of the analysis itself. In particular, a mathematical formalization of the fragmentation process is given in Section 2. After this, the peak interpretation problem is modeled by means of propositional logic. Each peak selected in the spectrum is used to build up a clause representing all the possible interpretations of that peak, as described in Section 3. A set of additional clauses, representing incompatibilities and other possibly known informations, is also generated. Then, we prove that all and only the coherent interpretations of the spectrum are given by the models of the generated set of clauses. Finally, for each of the above interpretations, all the possible compatible sequences of amino acids are generated, as explained in Section 4. Successful results on real-world peptide analysis problems are presented in Section 5.

2 A Mathematical View of the Fragmentation Process

When a peptide undergoes a MS/MS analysis, the occurring fragmentation process gives an essential support to the sequencing. 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 infor-

mation 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

$$A = \{a_1, a_2, \ldots, a_n\}, \qquad a_i \in R_+$$

of the weight values of such components that must be considered for the current analysis, together with the sets

$$Min = \{m_1, m_2, \ldots, m_n\},\ m_i \in Z_+$$
$$Max = \{M_1, M_2, \ldots, M_n\},\ M_i \geq m_i,\ M_i \in Z_+$$

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 $\delta \in R_+$ 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

$$Y^0 = \{y_1^0, y_2^0, \ldots, y_n^0\}, \qquad y_i^0 \in Z_+$$

of the numbers of molecules of each component (here the amino acids) contained in the overall compound (here the peptidic complex), and the number

$e_0 \geq 1$ of electrical charges retained by the ionized overall compound, the observed weight w_0 of the overall compound is given by the following equation,

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

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 2.1 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. Most of them are of standard types and will be here called canonical. In particular, there are three possible canonical N-terminal ionized fragments, called a-ion, b-ion, c-ion, and three possible canonical 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.

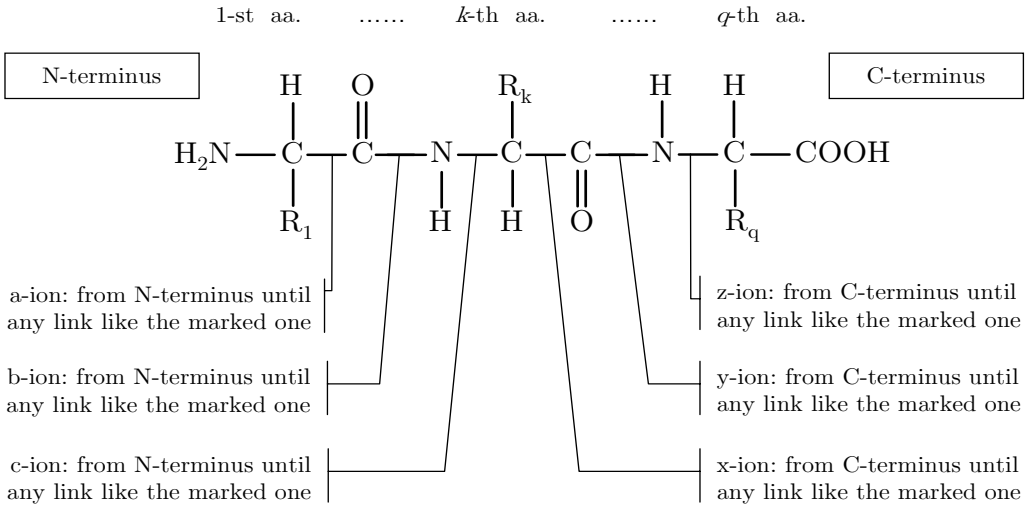


Fig. 2. Different types of fragments obtainable from a peptide.

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

$$W = \{w_1, w_2, \ldots, w_f\}, \qquad w_j \in R_+$$

of the weights corresponding to such peaks (so that w_0 remains the weight of the overall compound), the sets

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

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 canonical 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 $w_j, j = 1, \ldots, f$, and 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.

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

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 canonical) 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 2.2 When considering the spectrum reported in Fig. 1, and making the simplifying hypothesis of selecting only the numbered peaks (even if in practice a slightly larger set of peaks should be considered), we have $w_0 = 851.3$, $f = 9$, and $W = \{764.3, 651.3, 627.1, 538.2, 496.1, 425.1, 382.9, 201.0, 173.1\}$.

3 Clausal 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 3.1 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, in the sense 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 $w_j \to t, e$ 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

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

When considering for instance the 6 above canonical 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

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

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 3.2 Given a (real) single charge peptide of observed weight w_0, the *normalization* of such peptide produces a (theoretical) polymeric compound of weight $w_0 - (c_a + c_0)$, whose weight, as well as the weights of its fragments, is simply the sum of those of its components. Such normalization gives what is here called the *normalized peptide*. 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)

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

By 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 $Y^0 = \{y_1^0, y_2^0, \ldots, y_n^0\}$ of each amino acid contained in the real peptide of observed weight w_0.

Example 3.3 The normalized peptide corresponding to the real peptide of weight 572.69 of Example 2.1 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 have respectively the 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

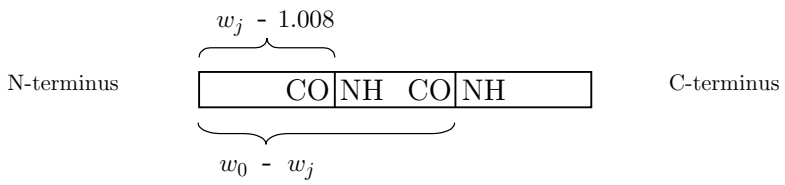
$$w_j - 1.008 \pm \delta$$

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 $w_j \approx 300$.

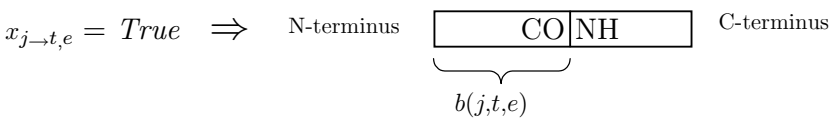


We now consider, for the each variable $x_{j \to t,e}$, with $j \in F,\ t \in T,\ e \in E$, the weight that the part of the normalized peptide going from the N-terminus to the cleavage corresponding to interpretation $w_j \to t, e$ would assume.

Definition 3.4 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 $b(j, t, e) \pm \delta$.



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 $(y_1, y_2, \ldots, y_n)^{tr} \in Z_+^n$ 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 $x_{j\to t,e}$, 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 3.5 Each model μ 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 μ 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 μ is an interpretation satisfying all the rules. Conversely, each interpretation satisfying all the rules corresponds to a truth assignment for the variables $x_{j\to t,e}$ 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 [14–17]. This problem is NP-complete [13] in the general case. However, for the average size of generated instances, solution times of a DPLL branching algorithm are very modest. 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. [18–21]). 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 [22], 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 [23] 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) [17], 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.

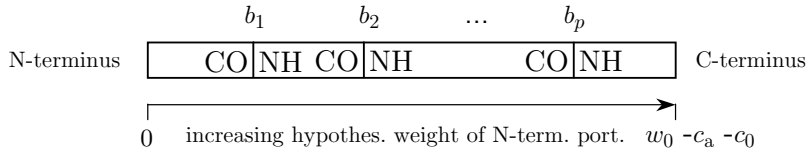
Example 3.6 When considering the compound of Example 2.2. ($w_0 = 851.3$, $f = 9$, and $W = \{764.3, 651.3, 627.1, 538.2, 496.1, 425.1, 382.9, 201.0, 173.1\}$), the possible components of Table 1, and allowing a-ion, b-ion, c-ion, x-ion, y-ion, z-ion, and double and single charges, we obtain a formula $\mathcal{F}$ with 108 variables and 4909 clauses, which has 3 models.

4 Generation of the Sequences Compatible with an Interpretation

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 μ for the generated formula $\mathcal{F}$, consider all the hypothesized weights of the N-terminal portions corresponding to all the *True* variables of μ. By ordering such values in increasing weight order, we obtain what we call the *succession of breakpoints* B^μ corresponding to model μ for the normalized peptide under analysis.

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

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



Define now a *gap* as the difference between two adjacent breakpoints (b_{h+1}, b_h), and a *subsequence* as the portion of the peptide spanning between two peptidic bonds corresponding to the two above adjacent breakpoints. Now we compute, for each value of gap $b_{h+1} - b_h$, all the non-negative integer vectors $(y_1, y_2, \ldots, y_n)^{tr} \in Z_+^n$ verifying the following equation.

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

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 δ instead of 2δ 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 specialized branching algorithm very closely related to DPLL SAT solvers, which proceeds by progressively fixing values for the y_i variables such that their domains $[m_i, M_i] \cap Z_+$ are respected (see Section 2), thus generating subproblems with progressively decreasing dimension. Internal nodes of the obtained search tree correspond to partial variable assignments, while the leaves correspond to complete variable assignments. Backtrack is performed when the weight corresponding to a partial variable assignment $\bar{y}$ exceeds the desired gap

$$\bar{y}_1(a_1 - c_a) + \bar{y}_2(a_2 - c_a) + \ldots + \bar{y}_n(a_n - c_a) - 2\delta > b_{h+1} - b_h$$

or when the weight of a complete variable assignment $\tilde{y}$ does not reach such gap

$$\tilde{y}_1(a_1 - c_a) + \tilde{y}_2(a_2 - c_a) + \ldots + \tilde{y}_n(a_n - c_a) + 2\delta < b_{h+1} - b_h$$

Such approach evidently has exponential time complexity. However, since each gap $b_{h+1} - b_h$ generally has a value corresponding to a very small number of components (never more than 4 or 5), the sets $S(b_{h+1} - b_h),\ h = 0, \ldots, p$ can be computed in extremely short times.

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 μ can be generated with the concatenation of such sets in all possible ways, operation which we denote by $\oplus$, but eliminating sequences violating the 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 4.1 When considering the formula $\mathcal{F}$ of Example 3.6 with 108 variables, 4909 clauses and 3 models, 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 numbered peaks, results are not as accurate as it would be possible when selecting more peaks.

5 Implementation and Computational Experience

The proposed approach is implemented in C++ and tested on a Pentium IV 1.7GHz PC. After the initial input routine, which (i) reads all informations about possible components and possible types of fragments and charges, (ii) reads the spectrum and extracts from it all peaks above a certain value, 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 [23], modified in order to search for all the models of the given formula. Then, for

each model μ of $\mathcal{F}$, the breakpoint succession is computed, and all the possible subsequences covering each gap are produced by means of a specialized branching algorithm and linked together. 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.

Input Data					Results				
w_0	f	$t_{\max}$	$e_{\max}$	n	x	v	r	$\mathcal{S}$	time
572.20	7	2	1	20	14	108	1	1	0.1
572.20	7	6	2	20	84	3571	2	2	1.9
851.30	18	2	1	20	36	543	1	4	0.5
851.30	18	4	2	24	144	6780	4	7	2.0
851.30	18	6	3	24	324	12642	10	16	5.6
859.12	20	3	1	40	60	2904	4	26	1.6
859.12	20	6	2	40	240	8156	5	29	4.1
913.30	16	2	1	20	32	539	2	7	1.0
913.30	16	6	3	20	288	10741	8	32	6.8
968.58	19	2	1	20	38	768	6	24	1.3
968.58	19	6	2	20	228	7021	10	38	4.1
1108.60	21	2	1	26	42	2687	8	18	3.5
1108.60	21	4	2	26	168	7456	16	64	12.2
1479.84	20	2	1	20	40	690	7	22	14.3
1479.84	20	6	2	20	240	8796	18	102	33.9
1570.60	22	2	1	21	44	2498	9	35	28.5
1570.60	22	6	2	21	264	9657	14	98	56.8
1607.69	27	2	2	26	108	5744	6	20	44.3
1607.69	27	6	3	26	486	22565	11	63	473.0

Table 2: Real-world peptide sequencing problems.

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. Those results are intended to give real-world examples of application, rather than exploring all the computational possibilities of the proposed procedure, since the latter is not the focus of present paper.

As observable from the table, results clearly depend on the choice of possible types and charges of fragments. This was of course expected. The number of sequences compatible with the given input data is sometimes large, but 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. Computational times are very moderate. 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.

6 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 means of a specialized branching algorithm very closely related to DPLL SAT solvers. Also due to the moderate dimension of the problems which the proposed approach generates, computational limits are completely overcome. The results of the reported tests on real-world peptide sequencing problems are very encouraging from the accuracy point of view.

Acknowledgements The author is grateful to prof. Gigi Gianfranceschi, from the University of Perugia, and to prof. Giorgio Koch, from the University of Roma "La Sapienza", for travelling together with him the long research path necessary for clarifying the various aspects of the fragmentation process described in Section 2, and to dr. Antonello Moscatelli for his precious contribution to the implementation work.

References

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

[2] R.L. Heinrikson. The Edman degradation in protein sequence analysis. In *Biochemical and Biophysical Studies of Proteins and Nucleic Acids*, T.B. Lo (editor) 285-302, Elsevier, New York (1984).

[3] 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).

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

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

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

[7] 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).

[8] 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).

[9] 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).

[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] 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).

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

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

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

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

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

[18] 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).

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

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

[21] 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).

[22] 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).

[23] 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).