# Mathematical Approaches to Polymer Sequence Analysis and Related Problems

Renato Bruni
Editor

# Mathematical Approaches to Polymer Sequence Analysis and Related Problems

Springer

*Editor*
Renato Bruni
Department of Computer and System Sciences
University of Roma "Sapienza"
Via Ariosto 25
00185 Roma
Italy
bruni@dis.uniroma1.it

ISBN 978-1-4419-6799-2 e-ISBN 978-1-4419-6800-5
DOI 10.1007/978-1-4419-6800-5
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010938717

# Preface

Many problems arising in biological, chemical, and medical research, which could not be solved in the past due to their dimension and complexity, are nowadays tackled by means of automatic elaboration. Powerful computers are indeed used intensively for solving many problems having biological origin, thus creating the emerging field of science called "bioinformatics." However, the success of such approaches depends not only on brute computational strength of those computers, but also, and often critically, on the mathematical quality of the models and of the algorithms underlying those solution procedures.

Solving a problem may be seen as converting information, in such a way that the solution of the problem (information in output) is extracted from its description (information in input), possibly passing through a number of intermediate states. By adopting this view, information handled when dealing with many of the above-mentioned problems becomes, at some stage, a sequence. Nature often encodes relevant information into sequences. Therefore, a central role in bioinformatics is played by sequence analysis problems or by the related problems of analyzing the effects or the behavior of some sequence.

The present volume offers a detailed overview of some of the most interesting mathematical approaches to sequence analysis and other sequence-related problems. Special emphasis is devoted to problems concerning the most relevant biopolymers (proteins and genetic sequences), but the exposition is not limited to them. A considerable effort has been made to render the volume comprehensible to researchers coming from either of the two hemispheres of bioinformatics: mathematics and computer science on one side, and biology, chemistry, and medicine on the other.

Rather than an exhaustive coverage of the topic, which would be clearly impossible to do in just one book, the volume is intended as a snapshot of the latest research development and of the potentialities that operations research and machine learning techniques bring in this interdisciplinary field of research. Moreover, the volume aims at bridging the two mentioned halves of bioinformatics that are still quite disjoint, promoting a cross-fertilization hopefully fostering future research in the field.

Primary selection criterion for the chapters has been scientific quality and importance. Additional selection criteria have been: (1) considering only approaches having a nontrivial mathematical basis; and (2) providing up to date contents not already largely available in other books published on similar subjects.

## Organization of the Volume

Due to the wide heterogeneity of the matter, from the point of view of both problems considered and techniques presented, it may be useful to the reader tracing the following short sketch of the volume organization.

The first part of the volume deals with problems originating from the study of protein sequences. Proteins and peptides are polymers made from units called amino acids, and a basic problem is the determination of their amino acid sequence when that is unknown. This is sometimes called analysis of the primary structure. In Chap. 1, Bruni deals with this problem, with a focus on peptides, since proteins are essentially polypeptide chains, and describes exact and complete approaches based on propositional logic.

To be able to perform their biological functions, proteins fold into specific spatial conformations. Another relevant problem is the determination of such structures, known as the problem of protein structure analysis or prediction. In particular, the disposition of highly regular substructures in the protein sequence, such as helices, sheets, and strands, is called the secondary structure, while the three-dimensional structure of a single protein molecule, and the spatial arrangement of the abovementioned elements of the secondary structure, is called the tertiary structure.

In Chap. 2, Di Lena et al. describe approaches to protein structure analysis based on decomposition, with specific attention to the secondary structure prediction and the protein contact map prediction by means of machine learning techniques. In Chap. 3, Patrizi et al. tackle again the problem of secondary structure prediction, performing a classification by means of nonlinear binary optimization techniques, with the aim of detecting isoform proteins considered as markers in oncology. Similarly, in Chap. 4, Biba et al. describe approaches to the protein folding prediction by modeling the sequence by means of Markov logic networks, that is, networks obtained by introducing probability in first-order logic.

The volume then gradually moves to problems originating from the study of genetic sequences. Deoxyribonucleic acid, or DNA, is a long polymer made from repeating units called nucleotides. It contains the genetic instructions used in the development and functioning of all known living organisms. In Chap. 5, Ceci et al. deal with the problem of discovering motifs, that are sequence patterns frequently appearing in DNA, RNA, or proteins, and therefore probably having specific biological functions. They are discovered by mining association rules in the three-dimensional space.

In Chap. 6, Mosca and Milanesi consider the problem of studying intermolecular interactions among DNA, RNA, and proteins obtained by means of sequence analysis techniques. When viewing those interactions at a system level, the dynamics of biochemical pathways can be simulated, and therefore better understood, by means of mathematical models.

In Chap. 7, Graça et al. deal with the problem of determining haplotype information, that is, genetic information inherited from ancestors, from genotype information, that is, all the genetic constitution of an individual, using approaches based on propositional logic. On related themes, in Chap. 8, Catanzaro describes the

problem of calculating phylogenies, that is, graphs representing the evolutionary relationships among species. Several optimization models for estimating them from molecular data such as DNA and RNA under different paradigms are explained and discussed.

In Chap. 9, Salvi et al. tackle the problem of performing studies of human genome by means of data mining techniques, known as genome-wide association studies, for a stratified population. This means that the individuals of the population are not uniform but carry different genetic backgrounds, and this often produces false association results. The effects of different statistical techniques are considered to devise an efficient strategy for overcoming this problem.

The last part of the volume considers problems originating from the study of polymers not having biological origin. Polymerization reactions can be divided into: (i) addition polymerization, producing the so-called addition polymers (also classified as chain-growth polymers, with some exceptions), which grow one monomer at a time, and (ii) condensation polymerization, producing the so-called condensation polymers (also classified as step-growth polymers), which grow eliminating small molecules during the synthesis. In Chap. 10, Montaudo deals with the problem of predicting the sequence distribution of addition polymers; while in Chap. 11, Montaudo discusses the same problem for condensation polymers, using in both a variety of mathematical techniques.

Rome, Italy
March 2010

Renato Bruni

# Contents

# Chapter 1
# Complete and Exact Peptide Sequence Analysis Based on Propositional Logic

**Renato Bruni**

**Abstract** Peptides are the short polymeric molecules constituting all the proteins. They are formed by the linking of amino acids, and the determination of the amino acid sequence of a peptide 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. This gives information about the molecular weight of the full peptidic molecule and of its fragments. Such information should be used in order to find the sequence, but this constitutes, in the general case, a difficult mathematical problem. After a brief overview of the approaches proposed in literature, and of their features and limits, the chapter describes in detail a promising one based on propositional logic. Differently from the others, this approach can be proved to be complete and exact.

## 1.1 Introduction

Peptides are short polymeric molecules formed by the linking of components called *amino acids* by means of covalent bonds called *peptide bonds*, in order to form a *chain*. Proteins are polypeptide chains; they are formed by a similar linking of amino acids, but the chain is generally longer. There are several different conventions to determine this distinction, see e.g. [4, 32].

The determination of the *sequence* of amino acids forming a peptide or a protein is one of the most important and frequent issues in many areas of chemistry, medicine and biology, as well as in several other applicative fields. In the case of peptides, this is often called de novo sequencing, whereas in the case of protein, this is often called determination of the primary structure. However, proteins are generally too extended for performing an accurate sequence analysis on the whole chain in a single step. Therefore, a protein molecule is usually divided into

R. Bruni (✉)
Department of Computer and System Sciences, University of Roma "Sapienza", Italy
e-mail: bruni@dis.uniroma1.it

R. Bruni (ed.), *Mathematical Approaches to Polymer Sequence Analysis and Related Problems*, DOI 10.1007/978-1-4419-6800-5_1,

its component peptides (via enzymatic digestion and subsequent fractionation with HPLC or capillary electrophoresis, [32]), and the original analysis is converted into a number of peptide analyses which are performed individually. It is worth noting that this problem has a theoretical structure that is able to represent various other problems of sequence analysis. At the very basic level, there is a set of possible components that are individually known a chain formed by some of such components, possibly repeated, whose sequence is not known and that cannot be inspected directly; and the aim is to determine this sequence of components forming the chain.

Nowadays, a widely used and well-established approach to peptide sequence analysis consists in the use of mass spectrometry [19, 20, 23, 28]. 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, especially when using the mass spectrometry/mass spectrometry methodology (also known as MS/MS, or tandem mass, [29]). 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 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 using several instrumental configurations, mainly triple quadrupole (QQQ), quadrupole time-of-flight (Q-TOF) and ion trap devices [20].

Since the weights of the possible components are known, and rules for determining the weights of sequences of known composition are available, the MS/MS information could be used in order to determine the unknown sequence of a peptide. This is, however, a difficult mathematical problem, as explained in detail in Sect. 1.2. Note that the presence of fragments constitutes the only source of information about the inner structure of the molecule under analysis: in the absence of fragmentation, the inner structure would be unknown. Several approaches to this problem have been proposed, as reported in Sect. 1.3. In particular, a promising approach [5] is based on a propositional logic modeling [12, 18, 31] of the problem, as explained in Sects. 1.4 and 1.5. 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. The off-line computation of the so-called weights database, which substantially speeds-up the sequencing operations, is described in Sect. 1.6. 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 illustrated by considering the case of peptides, but may be adapted to generic polymeric compounds submitted to mass spectrometry. Results on real-world problems, shown in Sect. 1.7, demonstrate the effectiveness of this approach.

## 1.2 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 Fig. 1.1. Further processing is then requested.

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 and 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 know the type of fragment that originated it (it could be either 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 [17]. However, when our compound is not in the databases (which may very well happen) or when the it 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 a univocal identification.

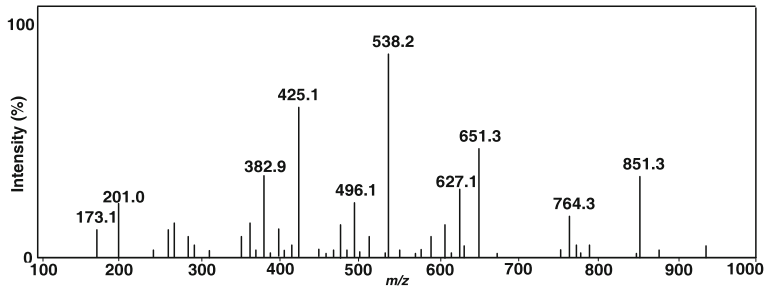



**Fig. 1.1** A MS/MS spectrum generated by collision-induced dissociation

**Definition 1.1.** 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 that have been selected as informative, but are instead due for instance to noise, non-standard fragmentation or spurious components. They are therefore not interpretable as standard fragments; hence, it may be the case that not even a sequence exists which is compatible with the given spectrum. In this case, the best that can be done, informally speaking, is being compatible with as many peaks as it is possible.

**Definition 1.2.** A sequence of components is $\nu$*-compatible* with a given spectrum if every informative peak in the spectrum, except a number $\nu$ of them, admits an interpretation as a standard fragment of that sequence. This number of uninterpreted peaks will be called the *mismatch* number $\nu$.

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

**Definition 1.4.** The *results* of a procedure are all the sequences that are given as the outcome of the analysis procedure.

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

**Definition 1.5.** 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 1.6.** 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.

A result given by an approximate procedure may just leave some informative peaks without an interpretation as standard fragments, or may give interpretation that

are not numerically precise. Note that this concept of approximate results is more general and less precise than that of $\nu$-compatible solution. Nevertheless, due to the stochastic aspects involved in the fragmentation process, these approximate results may sometimes be probable solutions.

Completeness and exactness are clearly positive features for a solution approach. However, complete and exact methods generally require larger computational times than incomplete or approximate ones [17, 21]. Note also that a complete and exact procedure correctly produces no results (or only results with mismatch $\nu > 0$) when the spectrum has no resolvents.

## 1.3 Different Approaches to the Problem

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 [24], Mass Seq [25], Peaks [26] and Spectrum Mill [27]. 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 different types mentioned. And, if the collision energy is increased, the peptide produces more fragments, but may also break at locations that are not the conjunction of amino acids, producing some non-standard fragments. Hence, every result given by the procedure is guaranteed to be a resolvent of the spectrum. On the contrary, there could be many resolvents of the spectrum not obtained as results because of the incompleteness of the series of fragments. The above approach should therefore be classified as heavily incomplete, though exact.

The second approach consists in iteratively generating, using Monte Carlo methods [8], 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 $20^{12} \approx 10^{15}$. Hence, 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 either 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 [15], 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 from considerable structural limitations.

Due to its combinatorial nature, the problem has also been recently approached by means of discrete mathematics. Specifically for the peptide sequencing problem, there have been, on one hand, the graph theoretical construction proposed in [14], which evolved into the dynamic programming algorithms proposed in [2, 11], and, on the other hand, the branching-based algorithm proposed in [7], which evolved into the propositional logic modeling proposed in [5]. The first approach has the advantage of requiring a computational time for finding each solution which is polynomial, hence tractable [15], 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 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; but can be improved as described in the rest of the chapter.

## 1.4 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. 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.

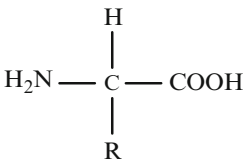


**Table 1.1** Commonly considered amino acids

| 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 | – |

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. Many of them cannot be specified in the genetic code; hence, the most commonly considered ones generally include the 20 reported in Table 1.1. Moreover, each amino acid may present one of the many possible modifications, such as phosphorylation, acetylation and methylation. 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 1.1 are just the average masses of the molecules.

An accurate and general 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 (in *unified atomic mass units* u, or daltons);
- Possible limitations on the position they can assume within a peptide chain;
- 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 a sequence analysis, the solution is obviously not known in advance. However, we often know which aspects of the solution can be considered as possible for the current analysis, and which ones cannot. For instance, we may know that a peptide under analysis contains at least a certain number of molecules of some amino acid or does not contain another amino acid, etc. At worst, if nothing else is known, simply every generic aspect of the solution should be considered as possible.

This may be formalized by evaluating 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

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

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 carboxyl 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, especially 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, an 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

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

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 peptide, the observed weight $w_0$ of the overall peptide is given by the following equation:

$$w_0 = \frac{\sum_{i \in N} (y_i^0 (a_i - c_a)) + c_a + c_0 e_0}{e_0} \pm \delta, \tag{1.1}$$

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 1.1.* 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 and c-ion, and three possible standard C-terminal ones, called x-ion, y-ion and z-ion, as illustrated in Fig. 1.2. 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 between the weight of each fragment and the weights of its components is the following.

$$w_j = \frac{\sum_{i \in N} [y_i^j (a_i - c_a)] + c_t + c_0 e_j}{e_j} \pm \delta, \qquad j = 1, \ldots, f \tag{1.2}$$

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 and 1.992 for z-ions.

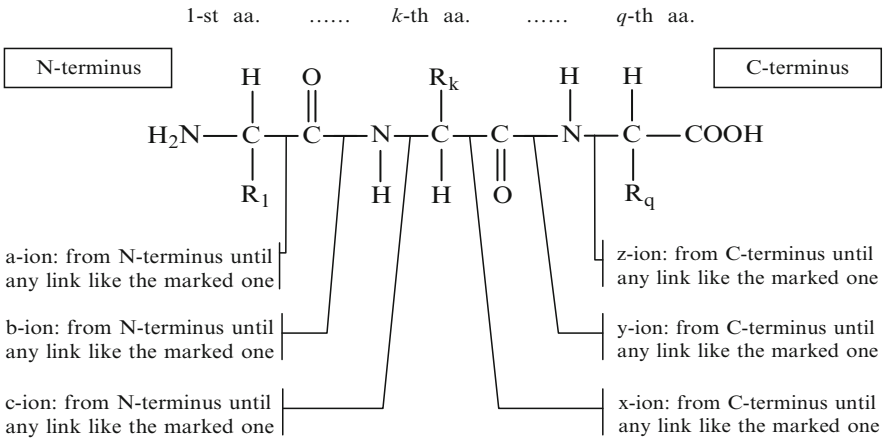



**Fig. 1.2** Different types of fragments obtainable from a peptide

In other words, the rules giving the weights of the six standard fragments having only one charge are as follows:

- a-ion weights the sum of its component amino acids, each of which decreased by 18.015, plus $(-28.002 + 1.008) = -26.994$;
- b-ion weights the sum of its component amino acids, each of which decreased by 18.015, plus $(0.000 + 1.008) = 1.008$;
- c-ion weights the sum of its component amino acids, each of which decreased by 18.015, plus $(17.031 + 1.008) = 18.039$;
- x-ion weights the sum of its component amino acids, each of which decreased by 18.015, plus $(44.009 + 1.008) = 45.017$;
- y-ion weights the sum of its component amino acids, each of which decreased by 18.015, plus $(18.015 + 1.008) = 19.023$;
- z-ion, finally, weights the sum of its component amino acids, each of which decreased by 18.015, plus $(1.992 + 1.008) = 3.000$.

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 1.2.* When considering the spectrum reported in Fig. 1.1 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 $W = \{764.3, 651.3, 627.1, 538.2, 496.1, 425.1, 382.9, 201.0, 173.1\}$.

## 1.5 A Logic Encoding of the Peak Interpretation Problem

Each peak of weight $w_j$ selected from the spectrum must be of one of the types $t \in T$ and have a charge $e_j \in E$, but the exact type and charge is in general 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 be originated by a certain amino acid sequence having a certain weight; if it is considered a b-ion, it cannot be originated by that sequence, but by another sequence having a different weight, 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.

**Definition 1.7.** The *peak interpretation problem* consists of 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*.

**Definition 1.8.** A set of interpretations for a set of peaks is coherent when all those interpretations respect a number of logical *rules* formalizing our knowledge of the problem.

Rules holding for every analysis are the incompatibility and multicharge rules, which are 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

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

When considering for instance the above six 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, which means: `peak` $w_j$ `is of type 1 and has charge 1 or it is of type 2 and has charge 1 or ... or it is of type 6 and has charge 2.`

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

Those clauses are called interpretation clauses. 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 1.9.** Given a (real) single charge peptide of observed weight $w_0$, the *normalized peptide* associated with it is a (theoretical) polymeric compound having 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 $Y^0 = \{y_1^0, y_2^0, \ldots, y_n^0\}$ of each amino acid contained in the real peptide of observed weight $w_0$.

*Example 1.3.* The normalized peptide corresponding to the real peptide of weight 572.69 of Example 1.1 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.

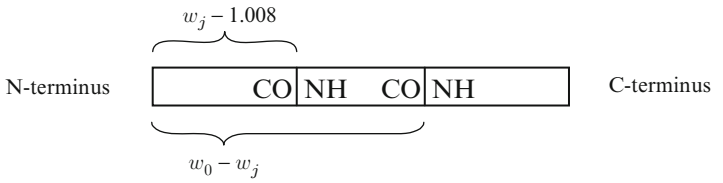
When a peak receives an interpretation, an hypothesis has been done about where the cleavage occurred in the peptide, and also about which was the chemical structure of the peptide in that point. Asserting that, for a single charge peptide of observed weight $w_0$, peak $w_j$ is, for instance, a single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

$$w_j - 1.008 \pm \delta$$

On the contrary, asserting that, for the same peptide, 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 an NH and a CO and that the part of such normalization going from the C-terminus to that cleavage has weight $w_j - 19.023 \pm \delta$. Therefore, the part of the same normalization going from the N-terminus to that cleavage weights

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

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 $w_j \approx 300$.

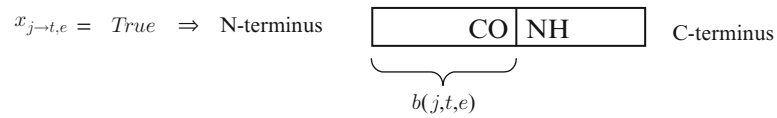



We now consider, for the each variable $x_{j \to t,e}$, with $j \in F$, $t \in T$, $e \in E$, the weight that the part of the normalized peptide going from the N terminus to the cleavage corresponding to interpretation $w_j \to t, e$ would assume.

**Definition 1.10.** 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 $x_{j \to t,e} = \mathit{True}$ 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 $(y_1, y_2, \ldots, y_n)^{tr} \in \mathbb{Z}_+^n$ verifying the following equation.

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

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 is 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 $\mathscr{F}$ in *conjunctive normal form* (CNF)

$$\mathscr{F} = C_1 \wedge C_2 \wedge \cdots \wedge C_v$$

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 $\mathscr{F}$ evaluates to *True* is known as a *model* of $\mathscr{F}$. We now have the following result.

**Theorem 1.1.** *Each model $\mu$ of the generated propositional formula $\mathscr{F}$ corresponds to 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 corresponds to a model $\mu$ of $\mathscr{F}$ can exist.*

*Proof.* The proof relies on the fact that the formula $\mathscr{F}$ contains by construction all the rules (peak assignment rules, incompatibility rules, multicharge rules) that a peak's interpretation must satisfy to be considered coherent. Therefore, each model

$\mu$ gives 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 $\mathscr{F}$ 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 [10, 16, 18, 30]. This problem is NP-complete [15] 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 [3, 9, 13, 22]. 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 coherent peaks interpretations, that is we are interested in finding all the models

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

of $\mathscr{F}$. 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.

*Example 1.4.* When considering the compound of Example 1.2 ($w_0 = 851.3$, $f = 9$, 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 1.1, and allowing a-ion, b-ion, c-ion, x-ion, y-ion, z-ion, and double and single charges, we obtain a formula $\mathscr{F}$ with 108 variables and 4909 clauses, which has three models.

In case $\mathscr{F}$ does not even have one model, this means that the considered sets of possible fragment types $T$ and/or possible charges $E$ are not enough to give an interpretation to all considered peaks. If $T$ and $E$ already include all possibilities that should be considered for the current analysis, they cannot be widened. In similar cases, the problem is originated by the presence of uninterpretable non-standard or noise peaks in the spectrum, which may be due to some experimental disturbance in the mass spectrometry analysis. For overcoming this type of problems, the mass spectrometry should be improved. When this option is not available, the formula $\mathscr{F}$ should be considered as an instance of the *maximum satisfiability* problem (Max-SAT) [30], which consists of finding a truth assignment for the variables $x_{j\to t,e}$ maximizing the number of clauses which evaluate to *True*. By doing so, some clauses will stay unsatisfied. Unsatisfied interpretation clauses correspond to a solution not interpreting some peaks, hence having a mismatch number $\nu > 0$. Unsatisfied incompatibility or multicharge clauses mean that not all rules for having a coherent interpretation can be respected in the current analysis. In any of these cases, the result of the analysis is less reliable, but the problem is in the input data.

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 (ILP). 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, [21]. 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 [10] for further details.

## 1.6 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 $\mathscr{F}$, 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.

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

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.