BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 17 2011, pages 2376–2383
doi:10.1093/bioinformatics/btr409

Structural bioinformatics Advance Access publication July 14, 2011

Elemental composition determination based on MS^n

Miguel Rojas-Chertó[1,2,*], Piotr T. Kasper[1,2], Egon L. Willighagen[1,3,4], Rob J. Vreeken[1,2], Thomas Hankemeier[1,2] and Theo H. Reijmers[1,2,*]

[1]Netherlands Metabolomics Centre, [2]Division of Analytical Biosciences, Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands, [3]Division of Molecular Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden and [4]Plant Research International, Wageningen UR, Wageningen, The Netherlands

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Identification of metabolites is essential for its use as biomarkers, for research in systems biology and for drug discovery. The first step before a structure can be elucidated is to determine its elemental composition. High-resolution mass spectrometry, which provides the exact mass, together with common constraint rules, for rejecting false proposed elemental compositions, cannot always provide one unique elemental composition solution.

Results: The Multistage Elemental Formula (MEF) tool is presented in this article to enable the correct assignment of elemental composition to compounds, their fragment ions and neutral losses that originate from the molecular ion by using multistage mass spectrometry (MS^n). The method provided by MEF reduces the list of predicted elemental compositions for each ion by analyzing the elemental compositions of its parent (precursor ion) and descendants (fragments). MS^n data of several metabolites were processed using the MEF tool to assign the correct elemental composition and validate the efficacy of the method. Especially, the link between the mass accuracy needed to generate one unique elemental composition and the topology of the MS^n tree (the width and the depth of the tree) was addressed. This method makes an important step toward semi-automatic *de novo* identification of metabolites using MS^n data.

Availability: Software available at:
http://abs.lacdr.gorlaeus.net/people/rojas-cherto

Contact: m.rojas@lacdr.leidenuniv.nl; t.reijmers@lacdr.leidenuniv.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 17, 2011; revised on June 23, 2011; accepted on July 4, 2011

1 INTRODUCTION

Metabolomics is the extensive examination of the metabolic phenotype, the metabolome, under a specific set of conditions. It has emerged as a functional key to understand the behavior of complex biological systems, including cells, tissues and biofluids. Currently, the essential challenges addressed in metabolomics are identification (full coverage of the metabolome with structural characterization) and quantification (accurately measure concentrations at low abundance levels). Metabolite identification to enable proper biological interpretation is the first step in the research of biomarkers, systems biology and drug discovery [Butcher *et al.* (2004); Dunn (2008); Kind and Fiehn (2010)].

Electrospray ionization (ESI) enabled mass spectrometry (MS) to become the most essential analytical tool used in both quantitative and qualitative metabolomic research (Cui *et al.*, 2000). In combination with liquid or gas chromatography, MS is the widely accepted standard method (Dunn and Ellis, 2005) for the analysis of metabolomic samples. In this way, each compound in the measured sample is characterized by a retention time and a mass value.

However, this characterization is not sufficient for metabolite identification, and the first step toward elucidation of the chemical structure from an unidentified compound is the determination of its elemental composition. An elemental composition expresses which and how many atoms constitute a particular chemical compound. Although each elemental composition has an unique molecular weight (mass), a molecular weight does not have a unique elemental composition. This situation is complicated by the instrumental precision, and the accuracy of the mass measurements limits this viability. Nowadays, mass spectrometry instrumentation such as time-of-flight, ion cyclotron resonance or magnetic sectors are able to measure mass-to-charge ratios (m/z) with a mass accuracy up to 1 ppm (parts per million) meaning that a mass of 100 Da is measured accurately up to four decimal places. In addition, modern desktop computers together with available software tools make it possible to generate and verify instantaneously all theoretical possible chemical elemental compositions for a given mass. Regrettably, several studies have shown that even high mass accuracy (smaller than 1 ppm) does not necessarily result in one unique assigned elemental composition (Kim *et al.*, 2006; Kind and Fiehn, 2006). The number of possible elemental compositions increases exponentially with the mass and the set of chemical elements taken into account. Generally, available methods that derive the elemental composition from a given mass use the following three steps: generation, filtering and matching. The generation step generates a candidate list with all possible elemental compositions enumerated systematically. The filtering step rejects all the elemental compositions that do not satisfy certain rules. The matching step compares the theoretical isotope patterns with the experimental one. The best match is reflected as the most probable elemental composition. Normally, the elemental composition annotation process starts when the user provides a constraint set consisting of the mass of the ion, the set of chemical elements to include and exclude, the limit range of number of atoms for each element and the mass tolerance (mass error window). There are two different approaches for generating the elemental composition and these depend on the

*To whom correspondence should be addressed.

search model used. First there is the deterministic search that enumerates all possible elemental compositions. This approach is computationally intensive but checks the complete solution space (Dromey and Foyster, 1980). Next to this approach is the local search approach where the investigated solution space is restricted. An example of this approach is developed by Zhang *et al.* (2005) who based the generation of possible elemental compositions on the optimization of the match between the theoretical and observed isotope patterns. Constraint rules are applied to limit the number of possible elemental composition candidates and they are generally based on empirical knowledge. They are well documented and their limitations have been analyzed elsewhere (Kind and Fiehn, 2007). Several examples of these chemical rules are the rings-plus-double-bonds equivalent (RDBE) (Dayringer and McLafferty, 1977) or double-bond equivalent (DBE), LEWIS and SENIOR rule (Senior, 1951) and the nitrogen rule. The application of heuristic rules (Kind and Fiehn, 2007) is a different approach to exclude non-valid elemental compositions. These constraints are derived from the analysis of chemical structure databases. The extracted rules heavily depend on the quality and diversity of the data in the database from which the rules were determined. As a consequence, those who do not have access to rich chemical structure databases will have limited success in extracting reliable rules. Fortunately, recent initiatives like Blue Obelisk movement (Guha *et al.*, 2006) or Science Commons are encouraging open source, open data and open standards that facilitate better access for scientists. Next to the mass, the observed isotope pattern is used to eliminate elemental compositions from the candidate list because different elemental compositions will have different mass spectral isotopomeric abundances. The isotope distribution provides information which is unique for a given elemental composition. Hence, the experimental isotope abundance pattern of a metabolite's mass spectrum can be compared with the theoretical one (Stoll *et al.*, 2006) to remove false candidates and produce a final list of results sorted according to the degrees of similarity, called the 'hit-list'. However, insufficient intensity, limited accuracy, overlapping isotope patterns and co-isolation complicate this approach and make extraction of the isotope pattern from the experimental data a difficult task. Furthermore, the peak intensities and masses sometimes are not accurate because often MS data is acquired using a malfunctioning centroiding algorithm (Erve *et al.*, 2009; Gu *et al.*, 2006). For these reasons, it is a challenge to search for additional rules which improve the efficacy of determining the elemental composition.

An new approach to introduce constraints in the search for the unique elemental composition is to use multistage mass spectrometry (MS^n) information. The technology used in multistage fragmentation mass spectrometry permits, by consecutive isolation and fragmentation of ions under low-energy collision-induced dissociation (CID), the creation of a set of hierarchical linked mass spectral data, as shown in Figure 1a. Whenever the number of ions is significantly large, each new generated fragment ion can be isolated and applied to new collisions. Through this procedure, a new mass spectrum of fragment ions is obtained. The hence created mass spectral tree data, with its interlinked ion relations, gives a richer description of the measured compound than the data obtained from single-stage MS or Tandem MS. The obtained tree topology more accurately characterizes the analyzed compound. It should be noted that the MS^n approach has not been developed to obtain an elemental composition of the molecular ion alone. The MS^n approach will become important for the identification of unknown metabolites if more MS^n data in reference databases will become available. The MS^n approach also suffers from similar issues as mentioned previously when using isotopic patterns (insufficient intensity, overlapping mass spectral trees) but especially when isotopic information is not available, because of limited accuracy, the MS^n approach may be of use.

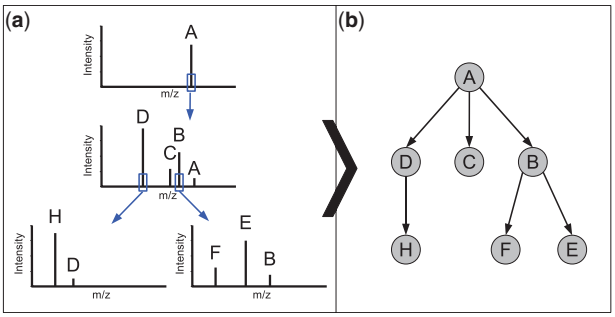

Fig. 1. Correlation between the spectral tree graph representation (**a**) and the fragmentation tree graph representation (**b**). In a spectral tree, the nodes are defined by spectra while in a fragmentation tree the nodes are characterized by ions and the edges by fragmentation paths.

Many approaches have been designed for metabolite identification using MS^n , but they often involve manual intervention by mass spectrometry experts and consequently are a very time consuming task (Cui *et al.*, 2000; Jarussophon *et al.*, 2009; Konishi *et al.*, 2007). Due to the complexity of processing experimental high-resolution MS^n data, it is to be preferred that for interpretation of this type of data a systematic computational process is used. For example, solving the elemental compositions for fragment ions is rather complex because for the fragment ions in MS^n spectra not the same constraint rules can be applied as for molecular ions. Nowadays more techniques are becoming available enabling automated processing of MS^n or MS/MS data like SmartFormula3D (Bruker Daltonics) (Tyrkkö *et al.*, 2010) and the method proposed by Rasche *et al.* (2011).

This article presents a new method to determine the elemental composition of a certain compound using MS^n data. It is shown that applying this approach to experimental MS^n data results in an unique elemental composition of the parent ion, their fragments and neutral losses, for several metabolites analyzed in our lab. Additionally, the dependency between mass accuracy and tree topology is analyzed and used to device guidelines for creating spectral trees with sufficient information to uniquely determine the molecular formula.

2 APPROACH

2.1 Algorithm constraining elemental composition generator

The multistage elemental formula (MEF) tool allows the determination of the elemental composition of ions and neutral losses from experimental MS^n data. First, the hierarchical mass spectral data need to be preprocessed and translated to a fragmentation tree representation. We define a fragmentation tree as a hierarchical organization of ions in a graphical form, such as shown in Figure 1. The tree in Figure 1 shows colored nodes that

define the ions/fragments, while the edges reflect the fragmentation reactions occurring. Fragments originating from a precursor ion are called child nodes and these are situated below the precursor ion/parent node in the tree. Each child node has only one parent. The so-called root node, shown at the top of the tree, has no parents and generally is the protonated or deprotonated molecular ion. All fragment ions that originate from the same parent ion with the same acquisition time, belong to the same experimental scan of the mass spectrum. We define the neutral losses as the residue of the fragmentation product which is not detectable by the mass spectrometer. We can calculate the mass of the neutral losses as the difference between the mass of the fragment and its precursor, if both masses are known. Note that the neutral loss masses do not have to correspond to one unique chemical structure. It could also express the sum of the masses for different neutral losses from consecutive fragmentations. The number of possible elemental compositions for a given mass depends heavily on the upper and lower limit of the number of atoms admitted to be present in the chemical formula. By narrowing the range of the number of atoms of each chemical element in the elemental composition, the list of theoretical possible elemental composition candidates decreases. The MEF tool, that we developed, is based on constraining for each chemical element the upper and lower limit (defined as the range) of number of atoms admitted to be present in the elemental composition. These ranges are derived from the elemental compositions of the precursor and fragments. Figure 3 shows an example of how an elemental composition range is generated from a list of elemental compositions. The range is extracted using the highest and lowest number of atoms in the complete elemental composition list. When a certain chemical element is not present in one theoretical possible elemental composition on the candidate list, the lower limit value is set to 0. When all elemental compositions in the list do not contain a certain element both the upper and lower limit values are set to 0. Repetition of this process for all possible precursor–fragment–child combinations produces new constraints that are used in the next cycle to further decrease the list of candidates. The procedure for solving the elemental composition of the precursor, fragment ions and neutral losses using fragmentation trees is summarized briefly in Figure 2. So the process begins by defining the input data for the MEF: the masses of all ions and the relations between them (the fragmentation pattern). The pre-processing paragraph in Section 3 describes in more detail how this information is extracted from the raw MS^n data. The first step, before any analysis can begin, is the correction of the ion masses. Depending on the detection mode used during acquisition (positive or negative mode), the masses of the ions must be adjusted (extracting or adding the mass of an electron) to make a valid comparison between the masses calculated from the theoretical-generated elemental compositions and the experimental mass. Next, an elemental composition generator generates and lists all possible theoretical elemental compositions for each ion and the neutral losses. The set of input parameters needed to generate the elemental compositions consists of the mass, mass accuracy (Δm), set of chemical elements to be present in the elemental composition and the range of the number of atoms for each element. If for some ion no chemical formula is derived, this particular ion and its fragments and the fragments of the fragments are removed from further analysis. For all ions and neutral losses, a candidate list of potential elemental compositions is generated, which is subsequently used

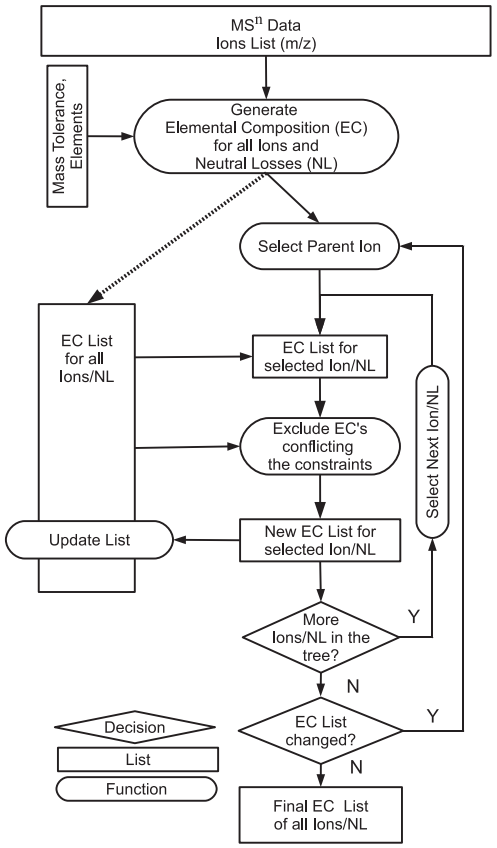

Fig. 2. Flow diagram of the MEF method extracting the elemental composition of all ions and neutral losses given MS^n data.

389.185179 +/- 5ppm		C	H	N	O	S	P
$C_{22}H_{23}N_5O_2$	$\Longrightarrow$	22	23	5	2	--	--
$C_{19}H_{26}N_4O_5$	$\Longrightarrow$	19	26	4	5	--	--
$C_{22}H_{32}N_1O_1S_2$	$\Longrightarrow$	22	32	1	1	2	--
$C_{19\cdot22}H_{23\cdot32}N_{1\cdot5}O_{1\cdot5}S_{0\cdot2}$	$\Longleftarrow$	$C_{19\cdot22}$	$H_{23\cdot32}$	$N_{1\cdot5}$	$O_{1\cdot5}$	$S_{0\cdot2}$	$P_{0\cdot0}$

Fig. 3. The example above shows how the elemental composition range is derived from an elemental composition list when a mass of 389.185179 Da. is provided with a mass tolerance of 5 ppm.

to extract the chemical formula ranges. These ranges correspond with the upper and lower number of atoms admitted for each chemical element in the candidate list (for an example, see Fig. 3). Once an elemental composition range has been obtained for the fragment ions and the neutral losses, three constraint rules are applied: the precursor consistency rule, fragment consistency rule and combinatorial consistency rule. These rules have the function to reduce the number of generated elemental compositions on each candidate by consulting the elemental compositions of the parent and the fragments.

The *precursor consistency rule* validates an elemental composition of a certain ion according to which elemental

compositions are assigned to its precursor ion. The precursor consistency rule defines that the number of atoms per element in the elemental composition of an ion cannot exceed the upper limit defined by the elemental composition range of the precursor ion. In other words, a fragment cannot contain more atoms of a certain chemical element than the precursor does.

The *fragment consistency rule* validates the elemental composition from the parent point of view. An elemental composition is considered valid when for all chemical elements, the number of atoms for a specific element is higher than the lower limit of the elemental composition range of the fragment(s). A precursor ion cannot contain less atoms of a certain chemical element than any of its fragments ions.

The *combinatorial consistency rule* uses the concept of conservation of mass. In a chemical reaction, the total mass of the reactants is equal to the total mass of the products. As fragmentation reactions are also chemical reactions, it is applied here as well not on the mass level but on the elemental composition level. The combinatorial consistency rule validates the elemental composition by checking if certain combinations of elemental compositions are present in the different candidate lists. There are three ways to do this.

- The elemental composition of a parent ion is accepted if at least once it is found as the sum of one elemental composition from a fragment and one elemental composition of the neutral loss. If no such combination is found, the elemental composition of the parent ion is removed from the list.
- The elemental composition of a fragment is accepted if at least once it is found as the difference of one elemental composition from a parent and one elemental composition of the neutral loss. If no such combination is observed, the elemental composition of the fragment is removed.
- The elemental composition of a neutral loss is accepted if at least once it is found as the difference of one elemental composition from a parent and one elemental composition of a fragment. If no such combination is found, the elemental composition of the neutral loss is removed.

These constraints are applied to each ion. When all ions are analyzed, the process is repeated for all ions starting from the parent node. This will be stopped when the list of candidate elemental compositions for all ions and neutral losses has not changed anymore. As soon as one or several elemental compositions are removed from the candidate list of a certain ion, this may result in removal of elemental compositions in neighboring ions. Table 1 displays an example of how the total number of possible elemental compositions is decreasing in each loop. These numbers are obtained after applying the MEF tool to MS^n data acquired for the compound Threonine. For simplification, only information of part of the nodes is shown. Any change to the list with potential elemental compositions will produce a new constraint in the next loop and will influence the lists of other ions. Finally, the iteration finishes when all ions conclude with the assignment of one unique elemental composition.

Table 1. Total number of elemental compositions obtained for each ion in each cycle during the application of the MEF tool for Threonine

Fragment			Cycle				
Ion ID	Precursor ID	Mass	First	Second	Third	Fourth	Fifth
1		120.065	30	30	10	2	1
2	1	102.055	6	6	4	2	1
3	2	56.049	3	3	3	2	1
4	2	84.044	7	7	6	2	1
5	1	74.060	19	13	4	2	1

The results were obtained using a mass tolerance of 10 ppm and the following set of atoms; C, H, N and O.

3 METHODS

3.1 Experimental section

A group of 12 compounds (see Supplementary Material about compounds) with known structure within a mass range of 150–450 Da was used to demonstrate the MEF method. The metabolites used in these experiments were purchased from Sigma (Sigma-Aldrich, Steinheim, Germany) and were of highest available purity. All compounds were dissolved at concentration of 0.1 mM. The solvent was 1:1 methanol:water (v/v) containing 0.1% formic acid. Solvents were of UPLC/MS quality and were purchased from Biosolve (Valkenswaard, The Netherlands). Mass spectra for these 12 compounds were obtained using a Finnigan LTQ-Orbitrap (Thermo Electron Corp.). The MS^n experiments were recorded using a data-dependent scanning function with the criteria to select the five most intense ions detected for MS^2 and the three most intense ions for the rest of the MS^n levels. For signal averaging, the mass spectrometer was set with five microscans. The Orbitrap was operated at 30 000 resolution, normalized collision energy of 35% and an isolation window of 1 Th.

3.2 Pre-processing

The information needed to start the MEF method consists of the exact masses, intensities for each peak, the specific acquisition times and the precursor scan. This information was extracted from the raw data using different existing external tools that were adopted to handle MS^n data. Ultimately, a pipeline for automated processing of multilevel mass spectral tree data was created by connecting these different tools. The raw MS data files (binary files) were converted to mzXML format (Pedrioli *et al.*, 2004) using ReadW software which is provided by the Institute for Systems Biology (the ISB) (see Supplementary Material for the mzXML files). We chose for the mzXML format because it is vendor independent and lists the precursor mass attributes that are used to find the relations between the different MS spectra (the hierarchical links of the spectral tree). The information about the relation between the different fragments of the spectral tree turned out to be not present in NetCDF (ASTM E2078-00 'Standard Guide for Analytical Data Interchange Protocol for Mass Spectrometric Data') files (another often used format to transfer mass spectral information from the analytical platform to data analysis software). For extraction of the ion peaks and finding relations between fragments of MS data, we used the freely available XCMS software (Smith *et al.*, 2006). XMCS reads mzXML files and identifies ion features (a specific *m/z* at a specific acquisition time and the precursor scan). With this information, it is possible to link specific fragment ion formation to the parent ion creating a hierarchical fragmentation path (fragmentation tree). The MS^n data were peak detected and noise reduced to exclude signals related to noise which could interfere in the analysis. The settings used to process all the MS data were the default XCMS settings. The final result is a table containing the information about the ion fragment peaks and their precursor ions which is used to initiate the MEF method for the extraction of the elemental compositions.

3.3 Data storage

It is important that after any information retrieval the outcome is stored for posterior handling. Here, the fragmentation pattern needs to be stored. There are several formats to store a chemical reaction representation, for example formats based on SMILES like mrv (Bode, 2004), the connection table-based formats, such as Symyx molfiles and rxn files, and the markup-based format, Chemical Markup Language (CML) (Murray-Rust *et al.*, 2001). At the moment, we generate fragmentation trees from MS^n data, which represents sequences of reactions. Each ion (reactant and product) is characterized by its elemental composition. Thus, we have chosen to use the CMLReact (Holliday *et al.*, 2006), an extension of CML, to connect the reaction components. CML has the ability to share all general XML features and unifies all available information for Internet publishing and computer processing. It supports various chemical concepts, such as molecules, reactions, spectra and other chemical data and data sources. The reaction is represented by molecular species behaving as reactants and products using the appropriate tags. In our application, we describe the ions only with their elemental composition, each product or reactant is defined with the tag elementalformula. When a fragment has more than one elemental candidate, these are put into a list.

3.3.1 Elemental formula generator code The generator for the automated extraction and handling of chemical formula given a molecular mass has been developed separately. It is available as part of the Chemistry Development Kit (Steinbeck *et al.*, 2003) (CDK) library. CDK is an open-source Java library for chemoinformatics and bioinformatics, which provides code for calculating QSAR descriptors, applying 2D and 3D modeling techniques, defining reaction mechanisms, etc. To generate elemental compositions first the mass accuracy, the set of the chemical elements and the maximal and minimal limit of the number of atoms to be present in the formula must be specified. The exact working of the algorithm generating all mathematical possible chemical element combinations is described in the article of Dromey and Foyster (1980). Furthermore, we integrated in the rCDK (Guha, 2007) package features to access certain functionalities needed for generating elemental compositions. The rCDK package provides an interface to CDK for R users, making the MEF method available in the R statistics software, for example, for direct integration with XCMS.

4 RESULTS AND DISCUSSION

The MEF method facilitates the analysis of MS^n data, which is not sufficiently explored in the mass spectrometry field yet. In a first series of experiments, we explore what mass accuracy is needed to resolve the elemental composition using MS^n data varying the set of chemical elements in the elemental composition and the MS^n level taken into account. Of our particular interest was to see to what extent the incorporation of the additional information in MS^n data, widens the ppm accuracy needed to uniquely identify the molecular formula corresponding to the measured mass.

To determine the needed accuracy value for obtaining a single elemental composition, the MEF method was repeatedly run with different values set as mass tolerance for all ions. The loop started with a rather large mass tolerance value of 180 ppm and we stopped decreasing it until the unique and correct elemental composition was found for the fragmented compound.

In the first experiment, MS^n data of 5-hydroxy-lysine was analyzed. 5-hydroxy-lysine was chosen because it does fragment in multiple high mass fragments and spectra can be acquired till MS level 5. With XCMS software, a peak list of 12 mass fragments was extracted. The MEF calculations were executed using four different sets of chemical elements: CHNO, CHNOS, CHNOSP and CHNOSPSi. We also included in the analyses Si, a heavier atom than C, N, P and S, to show the effect of the MEF tool of generating possible candidates when the atoms are more different in mass.

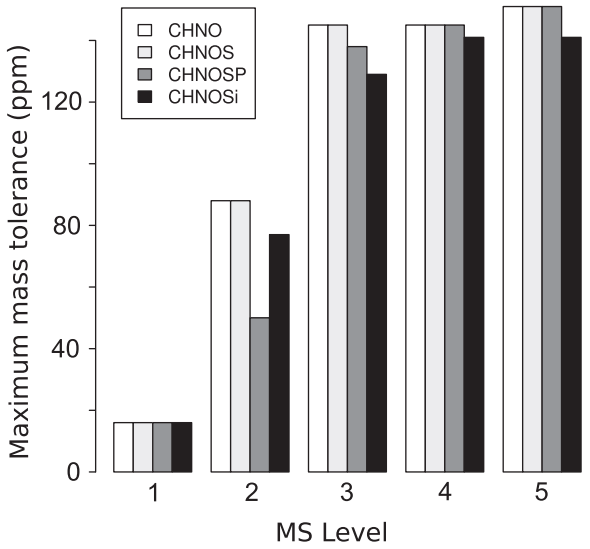

Fig. 4. The mass tolerance needed to generate one unique elemental composition for 5-hydroxy-lysine. Different sets of chemical elements and MS levels were taken into account. The range limit of atoms for each element are between 0 and 50 for C, N, O, S, P, Si, and 0–100 for H.

The results, see Figure 4, show that inclusion of additional MS^n levels in the extraction of the elemental composition has a high influence on the mass tolerance needed. For the fragments with low masses (found in the highest MS^n levels), a relative short list of candidate elemental compositions is generated with the consequence of putting stronger constraints on the precursor ions. As a consequence, the needed mass tolerance value to end up with one unique elemental composition can be higher (less accurate mass data are needed). There is a direct relation between the number of nodes (fragment ions) in the MS tree and the number of edges (fragmentation reactions). The number of nodes in a MS tree depends on the depth of the tree (the MS^n level) and the width of the tree (fragment ions on a certain MS level). More edges in the MS tree lead to more dependencies between elemental formulas list, ultimately leading to stronger constraints and a less stronger need for high accurate MS data. Another important factor influencing the outcome using the MEF tool is the set of different chemical elements to be included in the elemental composition calculations. Figure 4 shows that a higher mass accuracy is needed to assign one unique chemical formula when more different chemical elements are taken into account.

In the second experiment, mass spectral tree data was acquired and processed for 12 different metabolites, with masses between 150 and 450 Da, containing the following chemical elements: C, H, N, O, S and P. For all metabolites, MS spectra were acquired up to MS level 5. Each metabolite fragments in a different way resulting in an unique fragmentation tree topology for each of them. Again the efficacy of the MEF tool was tested. For all metabolites, the mass accuracy needed to determine the correct elemental composition of the parent ion was studied. Different metabolites with different molecular weights were compared while including different MS levels in the calculations. Figure 5 summarizes the results of

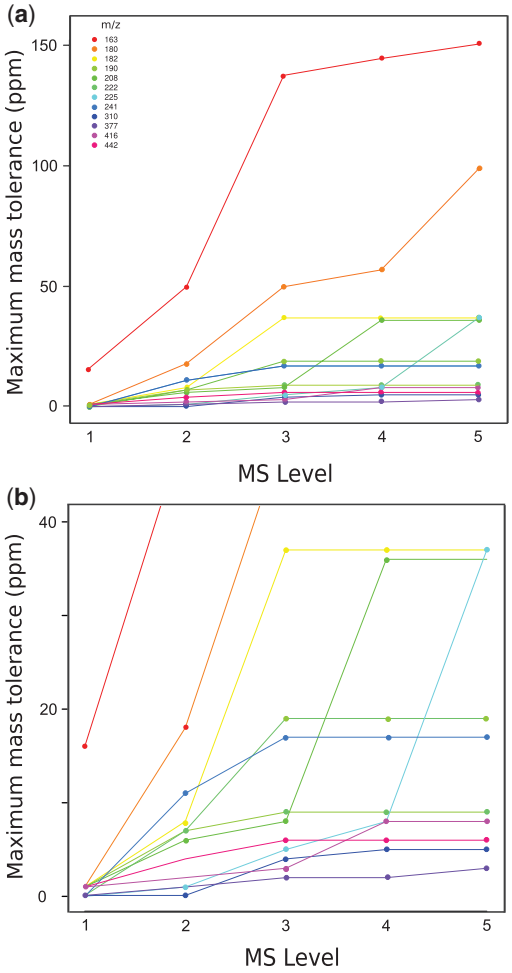

Fig. 5. The mass tolerance (mass error window) needed to obtain one unique elemental composition for 12 different metabolites. Different MS levels are taken into account. Figure (**b**) is an enlargement of figure (**a**). All masses or *m*/*z* given are those of the [M + H] ions.

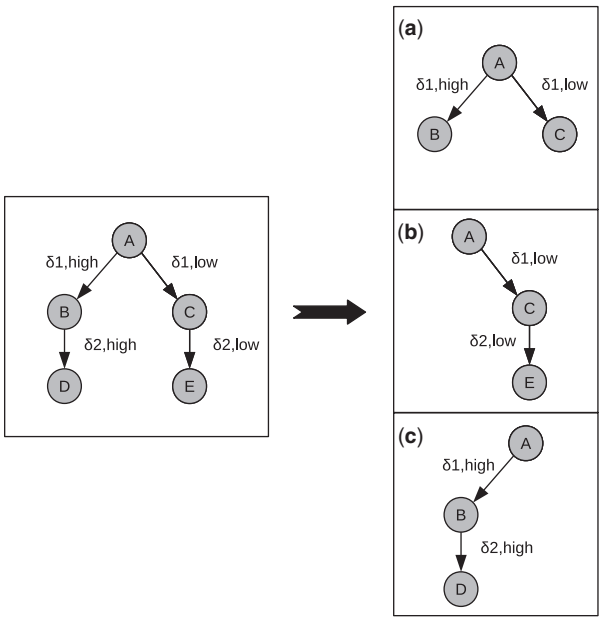

Fig. 6. Tree topology template used to check what the most informative nodes and edges are in an acquired fragmentation tree. Three possible scenarios to study are extracted. They represent the acquisition of fragmentation trees with: (**a**) wide/parallel mode, (**b**) linear/serial model with high masses and (**c**) linear/serial mode with low masses. The ions are represented by a circle. The δ values are the masses of the neutral losses calculated from the difference between the mass of the precursor ion and the fragment ion. The circle C and E are the fragments with the highest mass while B and D are the fragments with the lowest mass.

this experiment. When information from more MS^n levels (depth of the tree) was included in the determination of the elemental composition, all metabolites show that a less tight mass tolerance is needed. Again, taking advantage of the fact that more fragments are participating in the constraining process, the MS data has to be less accurate to determine the unique molecular formula. This approach shows that MS level is a relevant factor to help with the assignment of the elemental composition. From Figure 5, we can conclude that taking into account high MS levels allows us to set as a parameter higher mass tolerance error (or the instrument does need less mass accuracy) to determine the correct elemental composition using the MEF tool. To get one single elemental composition for the 12 different metabolites shown in Figure 5, on average MS^1 data of 1.83 ppm accuracy is needed while if MS^n data are available on average 35.58 ppm accuracy is needed. This effect is stronger for metabolites with a low molecular weight. Fragmentation of low weight metabolites result in fragments having relatively short elemental composition candidate lists ultimately leading to stronger constraints in the application of the MEF tool. The MEF tool accepts as input mzXML files, which makes the approach vendor and/or instrument independent. It can be applied both to high accurate instruments with limited fragmentation abilities (e.g. QTOF instruments limited to MS^2) and low accurate instruments using extensive MS levels [non-FTMS (ion trap) with MS^n].

The next exercise was executed to find out what the most informative edges in the fragmentation tree are for generating the molecular formula. Acquiring mass spectral tree data is time consuming and the aim of this exercise was to see which part of the tree gives the most relevant information for the MEF tool to ultimately guide data acquisition. As such, the results of this experiment would reduce the number of MS^n measurements to be taken. There are two approaches explored here to limit this number: first the acquisition of an MS^n tree can be tuned to generate deep or wide spectral trees depending on the experimental conditions set. The second acquisition approach is the selection of high or low ion masses for generation of the next fragment. To investigate both options, the tree topology template represented in Figure 6 was used. It consists of five nodes (representing the fragments), which can be extracted several times as subtree from all available fragmentation trees. Characteristic for this template is that the right side is formed by ions with high masses and the left side by ions with low masses. Using this template, we wanted to see which scenario needs less mass tolerance error to generate the correct elemental composition. The three scenarios selected were as follows: (a) wide/parallel mode, (b) linear/serial mode with high masses and (c) linear/serial mode with low masses. The δ values are the masses

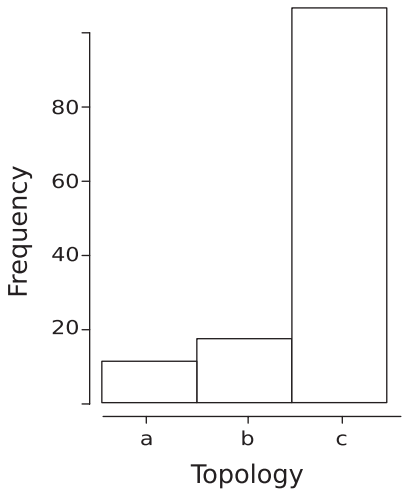

Fig. 7. A histogram showing the distribution of the best topology between the scenario (**a**), (**b**) and (**c**). The best topology is this that needs the lowest mass accuracy to generate one unique elemental composition.

of the neutral losses calculated from the difference between the masses of the precursor ion and the fragment ion. The template on the left side of Figure 6 was found 133 times in all fragmentation trees from the 12 metabolites.

Figure 7 shows in a histogram how many times a certain scenario was able to find the correct elemental composition using less accurate MS^n data than the other scenarios. Scenario (c) from Figure 6 clearly is the best data acquisition strategy to follow. In most cases, it is best to follow the linear approach and to go as deep as possible in the fragmentation tree. Furthermore, when there are multiple fragments available to choose from the fragmentation tree that contains the lowest masses is to be preferred.

In another exercise, Scenario (c) was further investigated. Two situations were considered where the δ 1 value was fixed and the δ 2 value divided into a high and low mass situation. Like previously, the best MEF results were obtained when the difference in mass between the precursor and fragment is as high as possible.

During the analysis of mass spectral tree data with the MEF tool, we encountered typical situations that lead to no assigned chemical formula for certain ions. In many cases, the selected ion turned out to be a false peak because of electrical or chemical noise or the mass tolerance applied to the ion was smaller than the experimental accuracy. When a valid elemental composition would be assigned to a false peak, this ultimately could lead to a false elemental composition assignment of the top parent ion. Although we did not encounter this in the data, we processed the next version of the MEF tool that contain features which allow the identification of these false peaks as well.

The runtime of the calculation does not only depend on the molecular mass of the compounds, but also on the number of fragments and neutral losses that are available in the spectral tree. The aim of the development of the MEF tool is a proof of concept to demonstrate the power of MS^n to determine the elemental composition of the molecular ion, fragment ions and neutral losses rather than an efficiently performing algorithm. The runtime is between seconds and minutes.

5 CONCLUSION

Due to fast recent instrumentation developments, MS^n has become a powerful tool for the characterization of metabolites. A new method was developed that can be applied for the processing and analysis of multistage mass spectrometry (MS^n) data. This method resolves the chemical elemental composition of a compound using constraints extracted from the predicted elemental composition of its fragments and requires a lower mass accuracy than conventional methods. The viability of the MEF method was tested with experiments on real MS^n data of several metabolites. The method does not only list the elemental composition of the parent ion, but also of its fragments and the neutral losses. The results presented here show that the method assigns very efficiently the correct elemental composition while reducing the needed accuracy to middle mass tolerance depending on the chemical structure of the metabolite and the topology of the fragmentation tree analyzed. For 5-hydroxylisine, the mass tolerance needed to solve the elemental composition jumps from 16 ppm to 150 ppm when additional fragmentation tree information is added as a constraint. This approach shows that the MS level is a relevant factor to help with the assignment of the elemental composition. If MS^n spectra are acquired to a higher level, the maximum mass tolerance to determine the correct elemental composition using the MEF tool is getting higher. To obtain the elemental composition of a protonated molecule (or adduct), this approach lowers the requirements with regards to mass accuracy of a mass spectrometer if MS^2 or higher MS level spectra can be obtained, and can be combined with the isotopic pattern of the protonated molecule (or adduct). This decreasing need for highly accurate data to solve the elemental composition by adding fragmentation tree information could help for those groups which cannot effort an expensive instrument with powerful resolution power. Alternatively, it will reduce the time needed to perform an identification. The output with the fragmentation pattern containing the elemental compositions is stored in a CML (Murray-Rust *et al.*, 2001) file format waiting for a proximate future for a standard exchange file format specific for metabolites. Currently, the MEF method provides a list of elemental composition candidates ordered according to the difference of the measured mass. In future work, it is planned to implement additional constraint rules into the method and also apply isotope abundance analysis to increase the identification accuracy. As we are able to characterize the fragments and neutral losses with the elemental composition, we are moving toward getting better understanding of the fragmentation patterns and use this information to identify the 'correct structure' of unknown compounds.

ACKNOWLEDGEMENT

We thank the editor and the anonymous reviewers for their helpful comments.

Funding: Research Programme of the Netherlands Metabolomics Centre (NMC) which is a part of The Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

Conflict of Interest: none declared.

REFERENCES

Bode,J.W. (2004) Reactor ChemAxon Ltd. *J. Am. Chem. Soc.*, **126**, 15317.

Butcher,E.C. *et al.* (2004) Systems biology in drug discovery. *Nat. Biotechnol.*, **22**, 1253–1259.

Cui,M. *et al.* (2000) Rapid identification of saponins in plant extracts by electrospray ionization multistage tandem mass spectrometry and liquid chromatography/tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, **14**, 1280–1286.

Dayringer,H.E. and McLafferty,F.W. (1977) Computer-aided interpretation of mass spectra. STIRS prediction of rings-plus-double-bonds values. *Org. Mass Spectrom.*, **12**, 53–54.

Dromey,R.G. and Foyster,G.T. (1980) Calculation of elemental compositions from high resolution mass spectral data. *Anal. Chem.*, **52**, 394–398.

Dunn,W.B. and Ellis,D.I. (2005) Metabolomics: current analytical platforms and methodologies. *Trends Analyt. Chem.*, **24**, 285–294.

Dunn,W.B. (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Phys. Biol.*, **5**, 011001.

Erve,J.C.L. *et al.* (2009) Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications for elemental composition determination. *J. Am. Soc. Mass Spectrom.*, **20**, 2058–2069.

Guha,R. *et al.* (2006) The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Informat. Model.*, **46**, 991–998.

Guha,R. (2007) Chemical Informatics Functionality in R. *J. Stat. Softw.*, **18**, 1–16.

Gu,M. *et al.* (2006) Accurate mass filtering of ion chromatograms for metabolite identification using a unit mass resolution liquid chromatography/mass spectrometry system. *Rapid Commun. Mass Spectrom.*, **20**, 764–770.

Holliday,G. *et al.* (2006) Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML vocabulary for chemical reactions. *J. Chem. Inf. Model*, **46**, 145–157.

Jarussophon,S. *et al.* (2009) Automated molecular formula determination by tandem mass spectrometry (MS/MS). *Analyst*, **134**, 690–700.

Kim,S. *et al.* (2006) Truly exact mass: elemental composition can be determined uniquely from molecular mass measurement at 0.1 mDa accuracy for molecules up to 500 Da. *Science*, **251**, 260–265.

Kind,T. and Fiehn,O. (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, **7**, 234.

Kind,T. and Fiehn,O. (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **8**, 105.

Kind,T. and Fiehn,O. (2010) Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.*, **2**, 23–60.

Konishi,Y. *et al.* (2007) Molecular formula analysis by an MS/MS/MS technique to expedite dereplication of natural products. *Anal. Chem.*, **79**, 1187–1197.

Murray-Rust,P. *et al.* (2001) Development of chemical markup language (CML) as a system for handling complex chemical content. *N. J. Chem.*, **25**, 618–634.

Pedrioli,P.G.A. *et al.* (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, **22**, 1459–1466.

Rasche,F. *et al.* (2011) Computing fragmentation trees from tandem mass spectrometry data. *Anal. Chem.*, **83**, 1243–1251.

Senior,J.K. (1951) Partitions and their representative graphs. *Am. J. Math.*, **73**, 663.

Smith,C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Steinbeck,C. *et al.* (2003) The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Informat. Comput. Sci.*, **43**, 493–500.

Stoll,N. *et al.* (2006) Isotope pattern evaluation for the reduction of elemental compositions assigned to high-resolution mass spectral data from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **17**, 1692–1699.

Tyrkkö,E. *et al.* (2010) Differentiation of structural isomers in a target drug database by LC/Q-TOFMS using fragmentation prediction. *Drug Test. Anal.*, **2**, 259–270.

Zhang,J. *et al.* (2005) Predicting molecular formulas of fragment ions with isotope patterns in tandem mass spectra. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* , **2**, 217–230.