Trends in Analytical Chemistry 69 (2015) 52–61

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

journal homepage: www.elsevier.com/locate/trac

Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics

Arpana Vaniya [a,b], Oliver Fiehn [b,c,*]

[a] Department of Chemistry, University of California Davis, One Shields Avenue, Davis, CA 95616, USA
[b] West Coast Metabolomics Center, Genome Center, University of California Davis, 451 Health Sciences Drive, Davis, CA 95616, USA
[c] Biochemistry Department, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Keywords:
Fragmentation tree
Ion tree
Mass spectral tree
Mass spectrometry
Metabolite identification
Metabolomics
MS^n
Multi-stage analysis
Tandem mass spectrometry
Unknown compound

ABSTRACT

Identification of unknown metabolites is the bottleneck in advancing metabolomics, leaving interpretation of metabolomics results ambiguous. The chemical diversity of metabolism is vast, making structure identification arduous and time consuming. Currently, comprehensive analysis of mass spectra in metabolomics is limited to library matching, but tandem mass spectral libraries are small compared to the large number of compounds found in the biosphere, including xenobiotics. Resolving this bottleneck requires richer data acquisition and better computational tools. Multi-stage mass spectrometry (MS^n) trees show promise to aid in this regard. Fragmentation trees explore the fragmentation process, generate fragmentation rules and aid in sub-structure identification, while mass spectral trees delineate the dependencies in multi-stage MS of collision-induced dissociations. This review covers advancements over the past 10 years as a tool for metabolite identification, including algorithms, software and databases used to build and to implement fragmentation trees and mass spectral annotations.

Contents

1. Introduction 52
2. Limitations of tandem mass spectrometry 53
3. Fragmentation trees and mass spectral trees 53
   3.1. MS^n ion tree for fragmentation analysis in natural products research 53
4. MS^n data-acquisition methods 55
5. Computational tools for MS^n and fragmentation trees 57
6. Tandem and MS^n spectral libraries and databases 59
7. Conclusions 60
   Authors' contribution 60
   Acknowledgements 60
   References 60

1. Introduction

Mass spectrometry (MS) is the dominant analytical technique in metabolomics. The elemental composition and structural information of a molecule can be readily determined by information provided by MS, such as accurate mass-to-charge ratio (m/z), isotope abundance [1] and fragmentation patterns [2]. The Metabolomics Standards Initiative (MSI) categorizes structure elucidation into four different levels: identification, annotation, characterization and classification [3,4]. These levels establish a thorough standard for the validation of metabolites that are identified across non-targeted metabolomic studies [4]. However, MSI does not provide a scoring schema to rank identified compounds within the identified and annotated categories, a caveat that was recently highlighted by metabolomics investigators [5]. Identification of metabolites refers to complete identification of the structure, including molecular connections and stereochemical assignments [6]. The identification process of small molecules in metabolomics is similar to that in other fields, such as toxicology and proteomics. All fields use accurate mass analysis, databases or libraries, and mass spectral fragmentations, such as LC-MS/MS. Some major differences between metabolomics and proteomics are the presence of multiply-charged ions from

* Corresponding author. Tel.: +1 530 754 8258.
E-mail address: ofiehn@ucdavis.edu (O. Fiehn).

http://dx.doi.org/10.1016/j.trac.2015.04.002
0165-9936/

peptides and the much larger chemical diversity in metabolomics and exposome analyses [7–9]. Synthesizing reference standards for confirmation of putative identifications is limited, time consuming, and uneconomical. According to MSI, annotation is putative compound identification in which the assignment of structures is highly likely, but not validated through chemical-reference standards [4]. Structure annotations are often ambiguous due to the large number of possible isomers, data inaccuracies, limited amounts of corroborating information, and human errors, including misclassification of sub-structures. However, annotation can also be viewed as a strategy to reduce the need for isolation of compounds and *de-novo* elucidation. The idea is to annotate mass spectra using the most probable elemental compositions found in public databases and to add additional orthogonal filters to decrease the number of structure hits [10].

Computer-assisted structural elucidation (CASE) encompasses structural dereplication using various analytical techniques from tandem MS (MS^2) and multi-stage MS (MS^n) to ultraviolet-visible (UV), infrared (IR) and nuclear magnetic resonance (NMR) spectroscopies. CASE first reduces chemical and spectral properties of an unknown compound, second generates candidate structures compatible with spectral features, and then ranks these candidates [11–13]. CASE can be used when manual interpretation of data is impractical and outcomes are unreliable using certain techniques, such as artificial intelligence, pattern recognition, library search, and spectral simulation [12,14]. Conversely, structural dereplication is performed by comparing experimental data to well-known databases that have standard reference data. Essentially, dereplication is a process to identify "known unknowns", which are compounds that are unknown at the time of detection and with further investigation are then found to be known compounds [15]. For example, the National Institute of Standards and Technology (NIST) database can be used to identify unknown compounds in gas chromatography-MS (GC-MS) studies [16]. Both structural dereplication and CASE are not considered *de-novo* identification because they rely on database searches with pre-existing known metabolites or reference standards [17]. Full *de-novo* identification by MS alone can hardly be achieved because isomers are difficult to distinguish by MS [10]. Mass spectral data inform about elemental compositions by combining accurate mass and isotopic information [1]. Collision-induced fragmentation data on the MS^2 or MS^n levels are used to find structural information from unique fragmentation patterns to test for the presence and the absence of functional groups. Interpretation of data in CASE may subsequently yield a partial structure or a sub-structure [12] (e.g., by using graphs that represent MS^n fragmentation-tree spectra in a hierarchical and data-dependent format). In CASE, rules, such as the calculation of "Rings plus double-bond equivalents" (RDBE), the nitrogen rule and the "even-electron rule", are applied when interpreting MS data to identify the formation of fragment ions and neutral species [18].

The scope of this review is to discuss advancements in techniques used by MS for structure elucidation, specifically the use of MS^n ion trees for small organic molecules with molecular weights less than 2 kDa.

2. Limitations of tandem mass spectrometry

While collision-induced dissociation (CID) MS/MS today is the dominant technique for library matching and interpreting fragment patterns to find structural information [6], using MS/MS alone falls short because product ions found in the MS/MS spectrum may be derived from intermediary ions instead of being produced directly from the molecular adduct precursor ion. For example, although epinine (deoxyepinephrine) conjugates in urine can be determined by MS/MS via precursor ion and neutral loss scans [19], MS/MS is unable to distinguish between positional isomers of such catecholamines. In addition, many fragment ions in MS/MS cannot be explained through fragmentation pathways even when structures are known [19]. Isomeric flavonoid O-diglycosides may yield different product-ion ratios in MS/MS fragmentation spectra [20]. However, such fragment-ion ratios cannot be used to infer interglycosidic linkages or glycan sequences in structural annotations of unknowns (Fig. 1) even though the authors successfully constructed a decision tree to differentiate these O-diglycosyl flavonoids [20].

Similarly, the annotation of positional sub-structures of taxanes in *Taxus* could not be achieved by MS/MS alone but only by using additional analytical methods [21]. Taken together, MS/MS certainly does not provide full structural information to elucidate an unknown compound completely. MS/MS fails to yield specific positional information of sub-structures, and many fragment ions remain unannotated with respect to presence of sub-structures or detailing fragmentation pathways.

3. Fragmentation trees and mass spectral trees

Trees are data structures defined by graph theory to organize and store data (e.g., the fragmentation process of an analyte of interest, or MS^n spectra generated by an ion-trap mass spectrometer). A tree is generated by nodes that are linked by edges (Fig. 2). Typically, the graphs are called fragmentation trees [23], family trees [24] or identification trees [25], if these trees show the fragmentation pathway of a molecule (Fig. 2A). Fragmentation trees are generated computationally to predict the fragmentation pathway of a molecule [23]. An implication of the fragmentation relationship between precursor ions and product ions is made before acquiring MS^n data. Conversely, ion trees or mass spectral trees refer to the sequential stages and relationships of mass spectral acquisition in MS^n processes, representing precursor and product ions as nodes and neutral losses as edges [26,27] (Fig. 2B). MS^n trees can therefore link ion-fragmentation pathways with (sub)structure relationships in a hierarchical order. An important aspect of MS^n trees is that they reveal both the dependency of precursor/product ion and product ion/product ion within the same MS^n stage or between different MS^n stages. This idea is rooted in the concept that any two MS^n spectra can ideally be treated as virtual MS/MS data: an ion has no memory. Hence, organizing large MS^n libraries will yield a tremendous expansion of publicly available MS/MS spectra, as long as each mass spectrum (Fig. 2B) is associated with a defined structure (Fig. 2A). For both fragmentation and mass spectral trees, computational methods are required to organize dependencies and extract specific information.

3.1. MS^n ion tree for fragmentation analysis in natural products research

MS^n multistage analysis provides means to link all product ions to specific precursor ions, hence enabling recursive reconstruction of fragmentation pathways that link specific sub-structures to complete molecular structures [28]. Oligosaccharides and sugar nucleotides were annotated using MS^4 ion trees with Mass Frontier 2.0 software [29], but the ion trap used lacked accurate mass capabilities to associate fragmentation rules unambiguously with potential fragmentation pathways to identify unknown metabolites detected in plant-phloem samples. Fabre et al. [30] successfully used MS^n to characterize structurally fragment ions and fragmentation mechanisms of flavonoid aglycones in negative-ion mode. MS^3 data supported fragmentation mechanisms, helped distinguish common neutral losses for specific sub-structures, and gave sufficient information to propose reasonable structures for fragments using both experimental and computational MS. However, for some