Trends in Analytical Chemistry 53 (2014) 41–48

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

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

Review

Computational mass spectrometry for small-molecule fragmentation

Franziska Hufsky [a,b], Kerstin Scheubert [a], Sebastian Böcker [a,*]

[a] Chair of Bioinformatics, Friedrich Schiller University, Ernst-Abbe-Platz 2, Jena, Germany
[b] Max Planck Institute for Chemical Ecology, Beutenberg Campus, Jena, Germany

ARTICLE INFO

Keywords:
Combinatorial fragmentation
Compound classification
Computational mass spectrometry
Computational method
Fragmentation tree
Library searching
Mass spectrometry (MS)
Rule-based prediction
Small-molecule fragmentation
Small-molecule identification

ABSTRACT

The identification of small molecules from mass spectrometry (MS) data remains a major challenge in the interpretation of MS data. Computational aspects of identifying small molecules range from searching a reference spectral library to the structural elucidation of an unknown. In this review, we concentrate on five important aspects of the computational analysis. We find that novel computational methods may overcome the boundaries of spectral libraries, by searching in the more comprehensive molecular structure databases, or not requiring any databases at all.

Contents

1. Introduction

Metabolomics covers detection, identification, and quantification of compounds of low molecular weight. Identification of metabolites poses a problem as, unlike proteins, these small molecules are usually not made up of building blocks, and the genomic sequence does not reveal information about their structure. Thus, a huge number of metabolites remain uncharacterized with respect to their structure and function [1].

Mass spectrometry (MS), typically coupled with chromatographic separation techniques, is a key analytical technology for high-throughput analysis of small molecules [1]. It is orders of magnitude more sensitive than nuclear magnetic resonance (NMR). Beyond information on the mass of the molecule, the compound can be fragmented and masses of the fragments recorded, revealing certain information about the structure of a compound. Several analytical techniques have been developed, where tandem MS is usually combined with liquid chromatography MS (LC-MS) [2], whereas gas chromatography MS (GC-MS) is coupled with electron impact (EI) fragmentation [3]. Given the huge amount of data produced in a high-throughput experiment, the manual interpretation of fragmentation spectra is time-intensive and often impractical [1]. So, an important aspect of small-molecule MS is the automated processing of the resulting fragmentation mass spectra.

Searching in libraries of reference spectra provides the most reliable source of identification. But this is only the case if the library contains a fragmentation spectrum from a reference compound measured on a similar instrument [4]. Unfortunately, spectral libraries are vastly incomplete. Recent approaches tend to replace searching in spectral libraries by searching in the more

* Corresponding author. Tel.: +49 36 41 94 64 50.
E-mail address: sebastian.boecker@uni-jena.de (S. Böcker).

0165-9936/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trac.2013.09.008

comprehensive molecular structure databases. Kind and Fiehn [5] give a survey of structure-elucidation techniques for small molecules using MS, whereas Scheubert et al. [6] review computational methods for this task.

In this review, we focus on the five basic approaches to dealing with metabolite fragmentation data, which are: (a) searching spectral libraries; (b) rule-based *in silico* fragmentation spectrum prediction; (c) mapping the fragmentation spectrum to the compound structure (combinatorial fragmentation); (d) predicting structural features and compound classes; and, (e) fragmentation trees (see Fig. 1).

2. Searching in spectral libraries

Given the fragmentation spectrum of an unknown metabolite, the straightforward approach to identifying the metabolite is looking up its fragmentation spectrum in a spectral library. For GC-MS, huge spectral reference libraries are routinely used; for LC-MS/MS, libraries contain fewer compounds and are limited in their availability. Database search requires a similarity or distance function for spectrum matching. Often, this is done using the "dot product" of the spectra. The spectra are treated as vectors $f = (f_1, \ldots, f_M)$ and $g = (g_1, \ldots, g_M)$, and the scalar product $\langle f, g\rangle = \Sigma_m f_m g_m$ is computed. This is particularly applied for unit mass accuracy data, where spectra can be directly mapped to vectors. For data with high mass accuracy, we can treat the spectra as continuous functions f, g with scalar product $\int f(m)g(m)dm$. Often, the raw peak shapes are not used but, instead, peaks are idealized as Gaussian functions. We can also introduce a weight function to weight the terms of the product differently, depending on the mass. Often, it is not the dot product that is reported but the enclosed angle θ or its cosine,

$$\cos\theta = \frac{\langle f, g\rangle}{\sqrt{\langle f, f\rangle}\sqrt{\langle g, g\rangle}}.$$

The spectral dot product is an advanced form of the most fundamental scoring, namely the "peak counting" family of measures that basically counts the number of matching peaks. Using the dot product for library searching is among the oldest computational techniques presented in this article, and has been developed independently of the task of searching for small compounds.

In 1994, Stein and Scott [7] evaluated the dot product against several other scoring systems, and found that it performed best of all. Several authors suggested modifications of the dot product, such as giving different confidence (weight) to different peaks; see [8,9] for two recent examples. Unfortunately, it appears to be a tough problem to outperform the basic dot product and its simplest modifications consistently and significantly.

The above scoring systems tell us which spectrum in the library best matches our query spectrum, and how to rank the remaining ones. But it cannot tell us whether this is a true or a bogus hit [10]. The reliable identification of a compound depends on the uniqueness of its spectrum. But the presence and the intensity of peaks across spectra are highly correlated, as these depend on the non-random distribution of molecular (sub-)structures. For example, benzene and fulvene have similar spectra, and a fulvene query spectrum would match a benzene database spectrum [10]. Hence, structurally-related compounds generally have similar mass spectra. This becomes a crucial problem when our database contains thousands of spectra. Unfortunately, little progress has been made in establishing the confidence of a compound identification using library search [11,12]. Citing Stein [10], the field of proteomics "has the luxury of being able to estimate 'false discovery rates' because of the ability to construct appropriate libraries of false identifications; such measures of reliability are not available for other classes of compounds". But we can also use the problem of similar spectra to our advantage: Since structurally-related compounds generally have similar mass spectra, false-positive hits may hint at correct "class identifications" if the true spectrum is not contained in the database [13]. Using fragmentation trees (see Section 6) as a detour in library searching allows us to compute such false-discovery rates (FDRs) for small-molecule MS.

The computational analysis of EI fragmentation spectra of small molecules via database search is generally simpler than for tandem MS data, as the fragmentation mechanisms are highly reproducible even across instruments, and reference spectra have been collected over many years [10]. However, LC-MS coupled with tandem MS fragmentation requires less sample preparation, and has other benefits, such as the known precursor mass of a compound. Fragmentation by tandem MS (such as collision-induced dissociation, CID) is less reproducible, in particular across different instrument types or even instruments [14]. Only first steps have been taken towards searching tandem MS spectral libraries [15], and these libraries are much smaller than those for GC-MS. Attempts have been made to create more reproducible, informative LC-MS fragmentation spectra [14,16,17].

For a comprehensive review on the fundamentals and difficulties of mass spectral libraries for compound identification, see Stein [10].

3. Rule-based fragmentation spectrum prediction

Spectral libraries are (and will always be) several orders of magnitude smaller than molecular structure databases. For example, PubChem currently contains about 30 million compounds, while even the biggest (commercial) spectral libraries, the National Institute of Standards and Technology (NIST) mass spectral library (version 11) and the Wiley Registry (9th edition) contain mass spectra for only 200 000 and 600 000 compounds, respectively. This gap may be filled by an accurate prediction of fragments (and their abundances) from the molecular structure of a compound. In this way, searching in spectral libraries can be replaced by searching in a database of theoretical mass spectra obtained from molecular structure databases. This trick has been very successfully used in proteomics for many years, as prediction of peptide fragmentation is comparatively easy.

To generate a set of candidate molecules, we can filter a molecular structure database using the molecular mass of the unknown, or even its molecular formula, if already known. However, we can use molecular structure generators to create a "private database", integrating further knowledge, such as substructure information.

Given a set of candidate molecular structures, spectra can be predicted by applying fragmentation rules to these structures, see Fig. 2. In principle, such rules can be learned from experimental data using data mining; but, until recently, experimental data were used solely to predict probabilities and, hence, intensities in the fragmentation spectrum [18,19]. In practice, these rules are manually curated from MS literature. First attempts at generating structural candidates and predicting their fragmentation mass spectra using general models of fragmentation, as well as class-specific fragmentation rules, were made as part of the DENDRAL project starting in 1965 [20,21]. However, the DENDRAL project failed in its major objective of automatic structure elucidation by mass spectral data, and research was discontinued [18]. Nowadays, there are three major commercial tools that predict MS fragmentation based on rules: *Mass Frontier* (HighChem, Ltd. Bratislava, Slovakia; versions after 5.0 available from Thermo Scientific, Waltham, USA), *ACD/MS Fragmenter* (Advanced Chemistry Labs, Toronto, Canada), and *MOLGEN-MS* [22,23].

Rule-based prediction systems were initially developed for prediction and interpretation of EI fragmentation data. EI spectra are