Trends in Analytical Chemistry 78 (2016) 23–35

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

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

Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects

Maria Vinaixa [a,b,c], Emma L. Schymanski [d], Steffen Neumann [e], Miriam Navarro [a,b,c], Reza M. Salek [f], Oscar Yanes [a,b,c,*]

[a] Centre for Omic Sciences, Universitat Rovira i Virgili, Avinguda Universitat 3, 43204 Reus, Spain
[b] Department of Electronic Engineering, Universitat Rovira i Virgili, Avinguda Paisos Catalans 26, 43007 Tarragona, Spain
[c] Spanish Biomedical Research Center in Diabetes and Associated Metabolic Disorders (CIBERDEM), Monforte de Lemos 3-5, 28029 Madrid, Spain
[d] Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dubendorf, Switzerland
[e] Dept. of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany
[f] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

ARTICLE INFO

Keywords:
Metabolomics
Databases
Mass spectrometry
Liquid chromatography
Gas chromatography
Identification
Mass spectral databases

ABSTRACT

At present, mass spectrometry (MS)-based metabolomics has been widely used to obtain new insights into human, plant, and microbial biochemistry; drug and biomarker discovery; nutrition research; and food control. Despite the high research interest, identifying and characterizing the structure of metabolites has become a major drawback for converting raw MS data into biological knowledge. Comprehensive and well-annotated MS-based spectral databases play a key role in serving this purpose via the formation of metabolite annotations. The main characteristics of the mass spectral databases currently used in MS-based metabolomics are reviewed in this study, underlining their advantages and limitations. In addition, the overlap of compounds with MS^n ($n \geq 2$) spectra from authentic chemical standards in most public and commercial databases has been calculated for the first time. Finally, future prospects of mass spectral databases are discussed in terms of the needs posed by novel applications and instrumental advancements.

© 2016 Elsevier B.V. All rights reserved.

Contents

* Corresponding author. Tel.: +34 977776617; Fax: +34 977301350.
E-mail address: oscar.yanes@urv.cat (O. Yanes).

http://dx.doi.org/10.1016/j.trac.2015.09.005
0165-9936/© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Metabolomics complements upstream biochemical information obtained from genes, transcripts, and proteins, widening current genomic reconstructions of metabolism and improving our understanding of cell biology, physiology, and medicine by linking cellular pathways to biological mechanism [1]. In order to achieve this, the following two technological platforms are most commonly used to identify and quantify metabolites: nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). They are often coupled to chromatographic techniques.

Unlike genes, transcripts, or proteins, which are biopolymers encoding information as a sequence of well-defined monomers, namely nucleotides and amino acids, metabolites are chemical entities that do not result from a residue-by-residue transfer of information within the cell. Instead, the extremely large diversity of metabolite structures in living organisms results directly from a series of chemical transformations catalyzed mainly by enzymes on environmental or dietary resources. Identifying and characterizing the structure of metabolites has become one of the major drawbacks for converting raw spectrometric data into biological knowledge, preventing metabolomics from evolving as fast as the other "omic" sciences [2–4].

Identification of metabolites is still evolving within the community, with active discussion on defining what constitutes a valid metabolite identification [5]. The Metabolomics Society, for instance, is currently assessing and developing an improved set of reporting standards [5–7]. The Chemical Analysis Working Group of the Metabolomics Standards Initiative (MSI; http://www.metabolomics-msi.org/) established four levels of identification so far [8]. Level 1 identification requires at least two orthogonal molecular properties of the putative metabolite to be confirmed with an authentic pure compound analyzed under identical analytical conditions. By contrast, for levels 2 and 3, the comparison against literature and database data is sufficient, and therefore rather than identifications, only annotations are achieved. Level 4 identification refers to unknown compounds. Following this classification, the majority of metabolites reported in metabolomics literature correspond to the annotation of known chemical structures described in databases. Most of these annotations are based on physicochemical properties (e.g., chromatographic retention time) and/or spectral similarity with public/commercial spectral databases [1]. This demonstrates the importance of having comprehensive and well-annotated MS-based spectral databases [9]. Conversely, relatively few studies comply with the standards of level 1 identification [1,10–12].

Proper usage and development of MS-based spectral databases, therefore, is essential for metabolomics to reach the status of other omic sciences [9]. Unfortunately, current databases are still far from containing experimental data of all known metabolites, despite attempts to increase and improve their content; one such example is the Metabolite Standards Synthesis Core (MSSC) initiative by the National Institutes of Health (NIH), aiming to generate new compound standards (http://www.metabolomicsworkbench.org/standards/nominatecompounds.php). The major limitation is the relatively small number of metabolites commercially available as pure standards, not to mention the large number of metabolites with unknown chemical structures that remain to be identified and characterized [12]. In addition, the transferability of mass spectral databases, particularly MS/MS, between MS instruments can impose some limitations, restricting the structural assignments of metabolites by empirically matching spectral values from pure standard compounds. Despite these limitations, the use of reference spectral databases is still one of the best approaches to annotate the structure of known metabolites when full isolation and structure determination by NMR or X-ray crystallography is not possible. Alternatively, novel computational tools that heuristically predict MS fragmentation patterns *in silico* have been developed to assist with identification of metabolites for which tandem MS data are not available yet in databases [13–19]. For electron ionization–mass spectrometry (EI–MS), it has been shown that fragmentation spectra can be simulated with quantum chemical and molecular dynamics methods [20], although the runtime is still too large (several thousand CPU hours per molecule) to simulate spectra for many compounds.

Freely accessible and/or commercially available compound databases currently used in the field of metabolomics provide information on chemical structures, physicochemical properties, spectral profiles, biological functions, and pathway mapping of metabolites. On the basis of these annotations, Fiehn et al. [21] classified these databases into two categories: (i) pathway-centric databases such as KEGG [22], Reactome [23], WikiPathways [24], or BioCyc [25] and (ii) compound-centric databases such as PubChem [26], ChemSpider [27], METLIN [28], MassBank [29], GMD [30], or Human Metabolome Database (HMDB) [31]. While PubChem, ChemSpider, and Chemical Abstracts Service (CAS) provide >60 million, > 35 million, and >100 million chemical compounds, respectively, they are not typically used in metabolomics because of the limited biological relevance of the vast majority of chemicals and the lack of mass spectral information. By contrast, some other compound-centric databases are also enriched with mass spectral information, which enables annotation of metabolites by matching mass spectral features of the unknown compounds to curated spectra of reference standards. Although these are much smaller repositories than PubChem or ChemSpider, mass spectral databases represent the first step in converting raw spectral data into metabolite annotations and thus biological knowledge.

This article aims at providing an overview of the state of the art on mass spectral databases most commonly used for metabolite annotation in metabolomics. Mass spectral databases and search engines aiming to assist identification of metabolites through spectral matching have existed for several decades for EI–MS, with other ionization methods developed over the past decade. In general, chromatographic and ionization techniques still determine the identification workflow and the most appropriate metabolite mass spectral reference database to be used. In subsequent sections, we will describe the latest versions of the most widely used mass spectral databases for liquid chromatography/mass spectrometry (LC/MS)- and gas chromatography/mass spectrometry (GC/MS)-based metabolomics. We focus on the advantages and limitations of these databases toward providing the annotation of metabolites.

2. Overview of the LC/MS-based untargeted metabolomics workflow

LC/MS-based untargeted metabolomics typically involves comparison of the relative abundances of metabolites in multiple samples without prior identification. By using liquid chromatography coupled to quadrupole time-of-flight [28] or Orbitrap-based mass spectrometers [32], hundreds to thousands of peaks with a unique mass-to-charge ratio *(m/z)* and retention time (*m/z*–RT pair) are routinely detected from biological samples in a profiling experiment. Each

Download English Version:

https://daneshyari.com/en/article/1249042

Download Persian Version:

https://daneshyari.com/article/1249042

Daneshyari.com