International Journal of Mass Spectrometry xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Mass Spectrometry

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

Mass spectral similarity for untargeted metabolomics data analysis of complex mixtures

Neha Garg [a], Clifford A. Kapono [b], Yan Wei Lim [c], Nobuhiro Koyama [a,d], Mark J.A. Vermeij [e], Douglas Conrad [f], Forest Rohwer [c], Pieter C. Dorrestein [a,b,g,*]

[a] Skaggs School of Pharmacy & Pharmaceutical Sciences, University of California at San Diego, La Jolla, California, USA
[b] Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California, USA
[c] Department of Biology, San Diego State University, San Diego, California, USA
[d] School of Pharmacy, Kitasato University, Tokyo, Japan
[e] CARMABI, Willemstad, Curaçao, Department of Aquatic Microbiology, University of Amsterdam, Amsterdam, the Netherlands
[f] Department of Medicine, University of California San Diego, La Jolla, California, USA
[g] Department of Pharmacology, University of California at San Diego, La Jolla, California, USA

ARTICLE INFO

Article history:
Received 2 April 2014
Received in revised form 30 May 2014
Accepted 4 June 2014
Available online xxx

Keywords:
Molecular networking
Mass spectrometry
Complex mixtures
Spectral matching
Cytoscape
Database search

ABSTRACT

While in nucleotide sequencing, the analysis of DNA from complex mixtures of organisms is common, this is not yet true for mass spectrometric data analysis of complex mixtures. The comparative analyses of mass spectrometry data of microbial communities at the molecular level is difficult to perform, especially in the context of a host. The challenge does not lie in generating the mass spectrometry data, rather much of the difficulty falls in the realm of how to derive relevant information from this data. The informatics based techniques to visualize and organize datasets are well established for metagenome sequencing; however, due to the scarcity of informatics strategies in mass spectrometry, it is currently difficult to cross correlate two very different mass spectrometry data sets from microbial communities and their hosts. We highlight that molecular networking can be used as an organizational tool of tandem mass spectrometry data, automated database search for rapid identification of metabolites, and as a workflow to manage and compare mass spectrometry data from complex mixtures of organisms. To demonstrate this platform, we show data analysis from hard corals and a human lung associated with cystic fibrosis.

1. Introduction

Study of complex mixtures at the molecular level using global untargeted metabolomics contributes a better understanding of the influence of various domains of life on each other, and their environment. Such mixtures may represent environmental samples taken from the soil or sewers; marine communities such as algae, corals, lichens, and diseased human organs such as the gut, oral cavity, lungs, pancreas, and kidneys, to name a few. Due to the enormous diversity of molecules present in such communities, the analysis of mass spectrometry data obtained from such samples often surpasses the capacity of modern informatics analysis. Modern mass spectrometry (MS) represents a powerful tool for studying metabolomics due to its speed, reproducibility, unsurpassed resolution, broad dynamic range, and ability to analyze samples of tremendous complexity [1–3]. Rapid development and advancement of mass spectrometry based techniques such as ultra high performance liquid chromatography (UPLC)–MS [4–6], nano LC–MS [7,8], nano-electrospray ionization-MS [9,10], nanospray desorption ionization-MS [11,12], LTQ and LTQ-Orbitrap hybrid Fourier transform ion cyclotron-MS [13,14], paper spray ionization-MS [15], direct analysis in real time-MS [16,17], and imaging-MS [18–20] have made it possible to generate tandem MS data on complex mixtures without a significant investment of time in regards to sample preparation. Although development of multiple high-throughput approaches such as metagenomics, metatrancritptomics, and metaproteomics have made it possible to begin understanding the physiology of complex mixtures, interest in metabolomics has seen a recent surge. Thus, tools for rapid acquisition of MS data are now available, but MS based analysis of complex mixtures is still plagued by the lack of robust tools for data visualization and metabolite identification which can then be used to derive correlations of this data with metagenomics, metatrancritptomics, metaproteomics and KEGG pathway

* Corresponding author at: Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, 9500 Gilman Drive, MC0751, La Jolla, CA 92093-0751, USA. Tel.: +1 858 534 6607; fax: +1 858 822 0041.
E-mail address: pdorrestein@ucsd.edu (P.C. Dorrestein).

http://dx.doi.org/10.1016/j.ijms.2014.06.005
1387-3806/

mapping analysis. The unmanageable amount of data, also referred to as "Big Data" generated using these approaches necessitate the development of methodologies to analyze and interpret the large volume of data as well as robust databases to classify and identify molecular features. This has been performed in the past by first prioritizing data using univariate and multivariate statistical analysis such as principal component analysis, partial least square discriminant analysis, *t*-tests, and hierarchical clustering which entails organizing data into matrices that are compatible with such analysis [21–25]. The peak picking and evaluation for this analysis is usually performed by using instrument specific softwares such as Bruker® Data Analysis and Bruker® Profile Analysis for Bruker, MassHunter for Agilent, MarkerLynx for Waters or publically available softwares such as XCMS [26], MZmine [27–29], and METIDEA [30] which can handle data from different instruments. Such analysis helps in organization and classification of data, highlighting important metabolites, intensities of which differ between sample types but does not aid in overall identification of individual molecules or their biological origin which is indeed the major bottleneck in metabolomics of complex mixtures. XCMS Online was developed to overcome this challenge. Herein, following the bioinformatic analysis of datasets, tandem MS data of peaks that differ significantly in intensities between two sample sets are matched with tandem MS data available in the database METLIN [31,32] to aid compound identification. A similar tandem MS mass spectral search approach has also been utilized in the past, for example, to identify metabolites in algal secretions [33] and for identification of lipids from LipidBlast database [34] using NIST MS Search GUI program. Tandem MS containing databases such as NIST [35], METLIN [31], HMDB [36], and MassBank [37] can be utilized for rapid annotation of Big Data generated from complex mixtures. Thus, retention time alignment and similarity in tandem MS spectra are now widely accepted characteristic metabolite features used to generate a high probable structural match. However, databases are not comprehensive in the diversity of molecules nor can they be used to begin annotating molecules that are simply related but are not present in the database. Furthermore, most molecules that can be found in complex mixtures of organisms have simply never been characterized before.

Thus, analysis of complex mixtures that yields millions of tandem MS spectra requires development of algorithms that organize Big Data in a fashion wherein one can exploit already existing data in the form of public databases or in-house generated databases to identify known molecules as well as new molecules. One such algorithm termed "molecular networking" exploits the chemical and structural similarity of molecules that is reflected in the similarity of the MS/MS spectra to organize large data sets [38–41]. Molecular networks is an effective approach that can be used to tease apart the origin and ID complex samples. Herein, similar MS/MS spectra are grouped into clusters where different clusters represent different classes of molecules. A network of such clusters can be easily visualized and analyzed using Cytoscape or other networking program. In Cytoscape, nodes corresponding to the database hits can be color coded and various relevant attributes such as retention time, parent mass, signal intensity etc. can be visualized in the Cytoscape or equivalent networking display software. Further, use of molecular networking allows rapid identification of molecules/metabolites that are similar to database hits since these molecules cluster together.

Herein, we show that molecular network analysis and automated matching of tandem MS data available in public/in-house database is one representative analysis workflow for molecular analysis of complex mixtures and to compare and contrast data sets of very different origin. In this publication, we demonstrate how one can tease apart complex mixtures at the molecular level by the use of molecular networking and database matching on a cystic fibrosis (CF)-afflicted human lung and hard corals both of which represent complex mixtures of microbial communities.

2. Results and discussion

2.1. Metabolomics workflow of complex mixtures

A typical workflow is shown in Fig. 1. The first step entails solvent extractions of the complex sample and its individual members. After extraction in appropriate solvents, UPLC–MS and tandem MS data is generated in second step. The Big Data generated in the second step is then organized using molecular networking via spectral matching of tandem MS spectra (Fig. 1a). The molecular networks generated via spectral matching create a network of spectral similarity and are visualized in Cytoscape. The similarity in fragmentation patterns due to similarity in structures under collision-induced dissociation dictates the clustering of metabolites in molecular networking. Each node consists of *n* number of identical MS/MS spectra present in the samples being analyzed (where *n* can be any number) and the neighboring nodes consists of MS/MS spectra that are related to each other. The clustering of MS/MS spectra within a node and between nodes is performed using an established algorithm namely "MS-cluster" [42]. The nodes are connected by edges where edge thickness is given by cosine score and depicts the relatedness between MS/MS spectra of connecting nodes. The mass spectrometry data analysis of complex mixtures generates thousands to millions of spectra that can generate very complex networks. Visualization in Cytoscape allows attribute mapping where the database hits, the hits from in-house generated pure compound libraries, as well as entire datasets of the individual components of the complex mixture are color coded to aid rapid identification (Fig. 1b). The neighboring nodes then represent molecules similar to the one obtained from database hit that were not present in the database, and can be dereplicated by comparing mass shifts in MS2 fragments [41]. Further, one can add entire data sets of different components of a complex mixture to tease apart the origin of these molecules. In order to demonstrate the potential of molecular networking in utilizing tandem MS data deposited in large databases and acquired on various different biological sample sets, we performed this analysis on anatomically distinct regions of CF-afflicted human lung and hard corals. Molecules with similar structures and hence similar fragmentation patterns clustered together and spectral matching with tandem MS library from NIST11 and METLIN aided identification of metabolites. Further, a combined molecular network of lung tissue and bacterial isolates obtained from the lung tissue was created to tease apart human and bacterial metabolites. The lung tissue and coral LC–MS/MS data was networked together to represent how metabolome of different complex communities can be compared and the commonalities identified using this approach.

2.2. Tandem MS guided molecular networking analysis

The sample set consisted of organic solvent extractions of ten anatomically distinct locations of ex-plant lung tissue and twenty *Pseudomonas* spp. isolates from a CF patient. In the current study, UPLC-ESI tandem MS data was collected on methanol and ethyl acetate extracts in the positive mode and organized using molecular networking (Fig. 2). The size of nodes represents intensity of the parent ion. As an example, one of the clusters is highlighted in Fig. 2b showing mass shifts of 2 Da, 14 Da and 28 Da between nodes suggesting a molecular family of fatty acids or lipids. The molecular networks form the lungs revealed matches to lipids, fatty acids, sterols, as well as drugs that were –administered

Please cite this article in press as: N. Garg, et al., Mass spectral similarity for untargeted metabolomics data analysis of complex mixtures, Int. J. Mass Spectrom. (2014), http://dx.doi.org/10.1016/j.ijms.2014.06.005