Marine Chemistry xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Marine Chemistry

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

Environmental metabolomics: Databases and tools for data analysis

Krista Longnecker [a,*], Joe Futrelle [b], Elizabeth Coburn [c], Melissa C. Kido Soule [a], Elizabeth B. Kujawinski [a]

[a] Woods Hole Oceanographic Institution, Marine Chemistry and Geochemistry, Woods Hole, MA 02543, USA
[b] Woods Hole Oceanographic Institution, Applied Ocean Physics & Engineering, Woods Hole, MA 02543, USA
[c] Independent Consultant

ARTICLE INFO

Article history:
Received 12 January 2015
Received in revised form 10 June 2015
Accepted 16 June 2015
Available online xxxx

Keywords:
Metabolomics
Data analysis
Database design

ABSTRACT

Metabolomics is the study of small molecules, or 'metabolites', that are the end products of biological processes. While -omics technologies such as genomics, transcriptomics, and proteomics measure the metabolic potential of organisms, metabolomics provides detailed information on the organic compounds produced during metabolism and found within cells and in the environment. Improvements in analytical techniques have expanded our understanding of metabolomics and developments in computational tools have made metabolomics data accessible to a broad segment of the scientific community. Yet, metabolomics methods have only been applied to a limited number of projects in the marine environment. Here, we review analysis techniques for mass spectrometry data and summarize the current state of metabolomics databases. We then describe a boutique database developed in our laboratory for efficient data analysis and selection of mass spectral targets for metabolite identification. The code to implement the database is freely available on GitHub (https://github.com/joefutrelle/domdb). Data organization and analysis are critical, but often under-appreciated, components of metabolomics research. Future advances in environmental metabolomics will take advantage of continued development of new tools that facilitate analysis of large metabolomics datasets.

1. Introduction

The exchange of organic compounds such as growth substrates, vitamins, and signaling molecules between microorganisms and their immediate surroundings is a central component of biogeochemical cycling in all environments. Genomics, transcriptomics, and proteomics data provide descriptions of how organisms may interact with these organic compounds. This information has led to key insights into the physiology of microorganisms and biochemical pathways that are potentially active in the marine environment. The field of metabolomics complements these data because it is used to directly assess active biochemical pathways, by measuring the end products of biological metabolic activity, i.e., the metabolites. Metabolomics studies can be grouped into two categories. Targeted metabolomics investigations obtain quantitative data on a pre-defined set of compounds, while untargeted metabolomics studies provide a broader exploration of metabolites with the goal of identifying new compounds (Patti et al., 2012). Environmental metabolomics is defined as the use of metabolomics techniques to characterize the metabolic response of organisms to natural and anthropogenic stressors in the environment (Viant, 2007).

Untargeted metabolomics datasets are large and multidimensional. While there have been improvements in computational programs that process mass spectrometry data files, tools are still needed to organize metabolites and their associated metadata to facilitate inter-experiment comparisons. Currently, two types of databases serve as repositories for information on organic compounds such as metabolites. The first type of database is focused on storage of chemical information of a compound, regardless of source. Two examples are the publically-accessible databases PubChem (Bolton et al., 2008) and ChemSpider (Pence and Williams, 2010) which contain vast amounts of data on organic molecules and can be searched by exact mass or compound name. On a smaller scale, METLIN (Zhu et al., 2013) is a curated database of biological compounds which allows searching by exact mass and by fragmentation spectrum. None of the above databases incorporate metadata and thus they do not provide an environmental context for any metabolite. In contrast, the second type of database allows contextualization of submitted experiments by explicitly including experimental metadata. The best example of such a database is the MetaboLights database (Haug et al., 2013; Steinbeck et al., 2012) which includes environmental data for the samples in the data repository. However, searches of MetaboLights currently access information about known compounds only. We are not aware of a database that allows searching of unknown compounds while retaining contextual metadata.

A metabolomics database has additional complications compared to other -omics databases. The structural and chemical composition of

* Corresponding author at: WHOI MS#4, Woods Hole, MA 02543, USA.
E-mail addresses: klongnecker@whoi.edu (K. Longnecker), jfutrelle@whoi.edu (J. Futrelle), eecoburn@gmail.com (E. Coburn), msoule@whoi.edu (M.C. Kido Soule), ekujawinski@whoi.edu (E.B. Kujawinski).

http://dx.doi.org/10.1016/j.marchem.2015.06.012

Table 1
Brief description of the sources of metabolomics data used to populate the boutique database. The samples span laboratory experiments ('lab exp.') and field expeditions, and include both intracellular and extracellular metabolite samples.

	Sample type	Extract type	# of samples	# of metabolites	Citation
Lab exp. #1	Laboratory experiment with *Thalassiosira pseudonana*	Extracellular metabolites	17	6047	Longnecker et al. (2015)
Lab exp. #2	Laboratory experiment with *Synechococcus elongatus*	Extracellular metabolites	16	10,158	Fiore et al. (in press)
Lab exp. #3	Laboratory experiment with *Ruegeria pomeroyi*	Intracellular metabolites	30	17,130	Johnson et al. (unpublished)
Lab exp. #4	Laboratory experiment with *Thalassiosira pseudonana*	Intracellular metabolites	24	4835	Kujawinski et al. (unpublished)
Lab exp. #5	Laboratory experiment with coastal seawater	Intracellular and extracellular metabolites	62	19,294	Liu et al. (unpublished)
Field #1	In situ samples, Pacific Ocean	Intracellular and extracellular metabolites	73	4236	(unpublished)
Field #2	Experiment with phytoplankton exudates, Atlantic Ocean	Intracellular and extracellular metabolites	27	7667	(unpublished)

genes and proteins are inherently simpler than that of metabolites because the number and diversity of building blocks are fewer. For example, gene sequences are comprised of only four or five possible nucleotides (A, G, C, T or U). Thus, a nucleic acid database such as GenBank (Benson et al., 2013) contains little chemical complexity and errors are primarily associated with interpretation such as gene annotation and homology assessments. In contrast, metabolites have no common building blocks, other than the elements of C, H, O, N, S and P; and their molecular structures and sizes are extremely diverse. Mass spectrometry-based metabolomics data are further complicated because each metabolite may be present as one or more adducts (e.g. $[M + Na]^+$ or $[M + H]^+$) with different mass-to-charge values. In addition, there is instrument-specific error associated with the mass-to-charge measurement. For liquid chromatography-based (LC) measurements, retention time varies as a function of chromatographic parameters such as column chemistry, mobile phase, and elution gradient. Finally, as with the gene-based databases, there is still the issue of identifying the metabolites and placing them into an environmental context.

An overarching goal of the research in our laboratory is the discovery, and subsequent quantification, of ecologically-relevant metabolites within marine ecosystems. For the database, we broadly define a metabolite as any organic compound observed in the marine environment. We use a combination of laboratory experiments and field sampling expeditions to uncover and to identify novel metabolites associated with important microorganisms in the marine environment. This goal requires the ability to store metabolomics data, to compare these data across different sampling scales, and to help focus time-consuming identification efforts on a tractable number of metabolites. As noted above, currently available databases cannot achieve these goals and thus we developed a boutique database for our laboratory. Inherent within this database development is a consideration of the computational challenges associated with the analysis of untargeted metabolomics data, in particular those data generated by ultrahigh resolution mass spectrometers coupled to a LC system. In this paper, we start with a review of freely-available and open source data analysis tools and databases for metabolomics data. We then describe our boutique database and conclude by providing examples of advances that rely on this joint consideration of field and laboratory samples.

2. Materials and methods

2.1. Designing a metabolomics database

Design of the boutique metabolomics database began with a series of meetings with domain scientists (here, the chemists), information scientists, and software developers. The purpose of these meetings was to establish the goals for developing the database and the desired outcome of the completed database. We employed 'use cases' to guide this process, and the outcome was an informal abstract information model and system design. Use case development is an integral methodology of the Tetherless World Constellation (Fox and McGuinness, 2008) and is an iterative method in which a small team of domain scientists and informaticists work together to rapidly develop prototype software to achieve the use case goal. The information model developed during this process captured semantic relationships between key concepts involved in the production and analysis of mass spectrometry data, as well as relationships central to extracting new knowledge from metabolomics experiments.

The information model and prototype system architecture were documented to serve as the initial phase of software prototypes. The prototype is a simple command-line interface on an object-relational model (ORM) implemented using Python and SQLAlchemy. These technologies, while not as powerful or scalable as technologies that would be appropriate for a larger database, have a number of features that make them attractive for prototyping. For example, the technologies are compatible with multiple platforms (e.g., Windows, Mac OS, Linux), are simple to install and configure, and enable rapid development, refinement and testing of new capabilities. The rapid prototyping

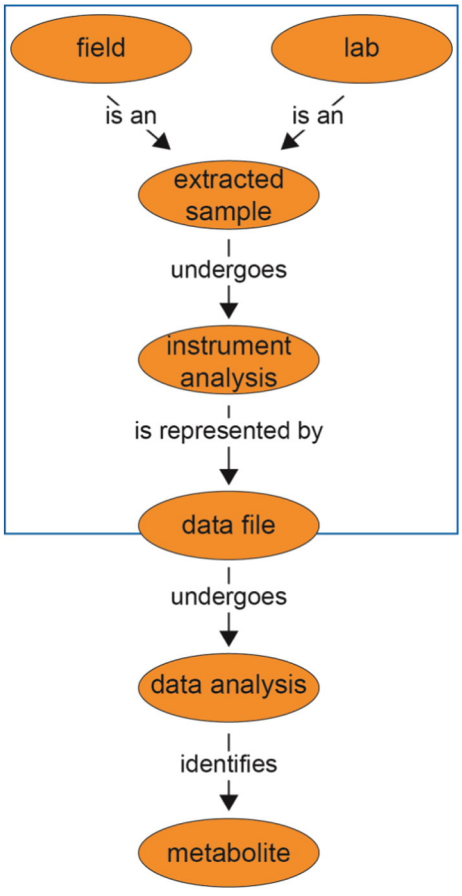

Fig. 1. A simplified version of the information model used to design the boutique database prototype. The complete information model is given in Fig. S1. The topics within the blue box are addressed in Kido Soule et al. (in revision). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)