Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-26 (2016) 286–288

Streamlining GC-MS metabolomic analysis using the M-IOLITE software suite

Christoniki Maga-Nteve * and Maria I. Klapa ******

**Foundation for Research & Technology, Institute of Chemical Engineering Sciences Patras, Greece (Tel:+30-2610-965-249 or -227, e-mail: xmaga@iceht.forth.gr; mklapa@iceht.forth.gr)*

***School of Medicine, University of Patras, Greece*

****Departments of Chemical & Biomolecular Engineering and Bioengineering, University of Maryland, College Park, MD 20742, USA*

Abstract: Metabolomics, as a rapidly growing omic analysis, has being used extensively to explore the dynamic response of biological systems in several diseases/disorders and contexts. Therefore, it has become commonplace in a wide variety of disciplines and there is an intense need for development of software suites that provide the user with a less complicated and invalid analysis. These suites must integrate meta-analysis, a standardized data normalization method and a safe repository for all types of biological samples. In the case of the Gas Chromatography-Mass Spectrometry (GC-MS) metabolomics, due to the complexity of the analysis, multiple procedures that are essential for the metabolite identification require special manipulation. Moreover, metabolomic analysis produces a vast amount of unidentified compound data, so there is a need for unknown peak identification methods. While a number of tools offer access to datasets, constantly providing new releases for data processing and the fact that considerable progress has been made in that area, there is no computational platform that emerges as a standardized approach which includes specialized normalization methods for GC-MS metabolomic analysis and incorporates the metabolic network analysis into data interpretation and unknown peak identification. To address these issues, as the datasets obtained from metabolomics experiments still remain extremely large and dense, we have implemented M-IOLITE, a computational suite for the efficient and automatic analysis of high-throughput metabolomic experiments. The aim of the suite is to streamline GC-MS metabolomic data analysis and to reduce complexity enabling the use of a friendly interface for processing, validating and annotating data. It integrates specialized normalization methods, a safe data repository and a peak library providing through its pipeline a useful tool which enables rapid and accurate analysis of the metabolomic profiles into an interactive system.

Keywords: Integrated suite, standardized data normalization method, repository, streamlining GC-MS metabolomic data analysis.

1. INTRODUCTION

Systems biology rapidly became vital due to the high-throughput technologies which enable the simultaneous measurement of hundreds of molecular compounds. High-throughput experiments can be conducted through various analytical techniques mainly Gas Chromatography - Mass Spectrometry (GC-MS), Liquid Chromatography - Mass Spectrometry (LC-MS) and the Nuclear Magnetic Resonance (NMR). In this study we focus on GC-MS metabolomics which aims to conduct the simultaneous determination and quantitative analysis of the small molecules (metabolites) that act as reactants or products in the metabolic reactions.

Mass spectrometry metabolomics (MS) workflow is a multistep procedure consisting of a variety of different modes in order to achieve a systems level perspective of metabolism and to reach a point of understanding or to discover biomarkers. Metabolomic analysis workflow that concerns analysis using MS is divided into analytical and computational parts. These two different parts are closely connected and are characterized by steps that arise as a natural consequence of every previous one.

Due to the huge amount and the complexity of the data that is produced from those experiments, there is need for software packages that provide efficient analysis and automatic data storage, as the automation of the metabolomic workflow can contribute to the reproducibility of data analysis and the minimization of error, while increasing the efficiency of the high-throughput analysis. In most cases, these packages do not contain every step of the computational part of metabolomic data processing.

Currently three of the most well known and publicly available software for MS metabolomics which are at the cutting edge of the field, are the Metabolights (http://www.ebi.ac.uk/metabolights/) which is mainly a data repository, XC-MS (https://xcmsonline.scripps.edu/) which mainly refers to LC-MS analysis providing peak identification, quantification and Metaboanalyst (http://www.metaboanalyst.ca/) which refers to multiple platforms and also provides options like identification and normalization. But there is no computational platform that includes specialized normalization and validation methods (Kanani and Klapa, 2007, Kanani, Chrysanthopoulos and

10.1016/j.ifacol.2016.12.140

Klapa, 2008) for GC-MS metabolomics and which incorporates the metabolic network analysis into data interpretation and unknown peak identification. Taking into consideration the needs of the GC-MS metabolomic data analysis, we developed M-IOLITE (Fig. 1.).

2. MATERIALS AND METHODS

2.1 Computational suite and repository peak library

This specific tool is designed for the streamlining of the metabolomic data analysis and also to reduce the complexity and the time required for these, through user-friendly environments and processes. The purpose is to produce efficiency and to maximize metabolic analyses by providing an optimal and organized procedure. Thus it integrates:

a) the capabilities of a standardized public repository

b) specialized normalization and filtering methods

c) metabolic network analysis for data interpretation and unknown peak identification

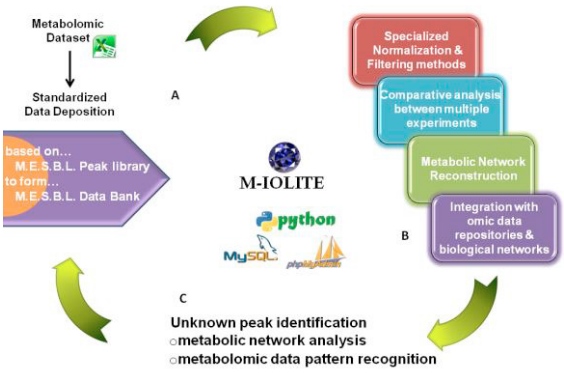

Fig. 1. M-IOLITE workflow divided into three parts. A) Metabolic data repository B) Metabolic data processing and interpretation C) Unknown peak identification

M-IOLITE is a Microsoft Windows based computational suite for high-throughput metabolomic data analysis which will be provided free to academic users upon request (miolite.iceht.forth.gr:8080) and it can be found in an executable format. It has been developed using programming language Python 2.7 (https://www.python.org/) and its available libraries that are provided through (https://www.python.org/) for data import, export and processing. Also, we used Mysql (http://www.mysql.com/) for the development of a peak library that comprises the identified peaks of the various experiments that have been conducted in our laboratory. The coordination of the database is carried out through phpMyAdmin v.4.3.11 (http://www.phpmyadmin.net/home_page/index.php) and wxFormBuilder v. 3.5 was used to design the graphical environment of the platform (GUI) (http://sourceforge.net/projects/wxformbuilder/). Through the software workflow the streamlining and the complexity reduction of the metabolomic analysis by developing a user friendly interface is provided for processing, validating and annotating metabolomics data. Our suite is divided into three parts which can run independently or not and is focusing on the GC-MS metabolomics.

The first part is associated with a standardized data repository based on the cross-validating in-house metabolic profile peak library of standard compounds and experiments that have been conducted in our lab, to form a metabolomic data bank.

The second part is related to specialized GC-MS metabolomic data normalization, filtering and quality control methods (Kanani et al. 2008; Kanani & Klapa 2007). In the case of GC-MS metabolomics, metabolites need to be converted into their stable derivative form in order to be detected. Some metabolites form more than one derivative in serial production over the course of its derivatization, like amino acids. Thus, appropriate normalization is required to avoid biases (Kanani et al. 2008; Kanani & Klapa 2007). Currently, our research group has the only normalization method for the GC-MS metabolomics.

The third part points to the use of metabolic network analysis and metabolomic data pattern recognition for unknown peak identification. In GC-MS, when a molecule leaves the chromatographic column, it gets bombarded by an electron beam and fragmented. Its fragmentation pattern along with the retention time in the chromatographic column can contribute to its identification. Thus, the tools must provide specialized peak identification methods based on the metabolite fragmentation patterns, as multiple peaks are identified as unknown ones. In this study, we will focus on the third part and in the pipeline we pursue in order to identify unknown peaks.

2.2 Unknown peak identification

For unknown peak identification, we incorporate information from different biological databases. More specifically, from the:

a) Golm Metabolome Database (GMD) (http://gmd.mpimp-golm.mpg.de/), which refers to the multiple trimethylsilyl (TMS) and methoxime (MEOX) derivative(s). Also, we extract information about the fragmentation pattern, retention time, molecular weight and many other related data.

b) KEGG (http://www.genome.jp/kegg/), which is a metabolic database from which relevant information about the metabolites and metabolic pathways can be retrieved.

c) Commercial NIST (version 2.0) (http://www.nist.gov/).

d) In-house library based on multiple experiments (Plants, Animals (Mice), Blood (plasma or serum), Cells (Bhk-HeLa) and standard compounds.

Apart from, the above databases we also download the reconstructed human network, Virtual Metabolic Human (VMH) (https://vmh.uni.lu/), in order to focus on molecules and reactions that are identified in humans. There are about 10 reconstructed networks in different organisms but initially we have started studying humans.

After connecting the GOLM, KEGG and VMH databases we incorporate this information in our database in order for it to be further used for the unknown peak identification. In particular, we match the multiple IDs like KEGG IDs, CAS numbers and EC numbers which are unique for every metabolite and reaction, to achieve an association between them. Then we filter out some information as unreliable data such as metabolite IDs that do not exist anymore or isotopes of the same fragment or ions derived from derivatizing agents etc. To address this issue, we edit a python script (Fig. 2.) where the user can provide the retention time in the chromatographic column and the fragmentation pattern of the unknown peak that wants to be identified. Then the script searches the database combination and if a match exists then a success rate is presented as well as a relative human id, in the case of the reconstructed human network. In order for it to be more precise in our method, we take into consideration the physicochemical properties of the metabolites like the molecular weight and specific chemical groups. So, some more filtering rules were added in our script, in an attempt to exclude data that would confound our results (i.e molecules with high molecular weight cannot be detected using GC-MS). Additionally, aiming to get more information we also recommend the commercial database NIST (version 2.0) and our peak library.

Furthermore, we incorporate the metabolic network analysis through the data pattern recognition to contribute to the identification resulting in a real biological meaning. In this case, we integrate multiple data from different samples and experiments, to end in a more complete, annotated and thorough dataset. Then we analyze the dataset through the Pavlidis Template Matching (Pavlidis & Noble 2001), in order to study the multiple metabolite patterns of known and unknown peaks, through the different experiments. If the pattern of the unknown peak matches with the pattern of known metabolites which take place in the same metabolic pathway then we try at least to chemically categorize it based on (Kanani & Klapa 2007), if not to identify them fully.

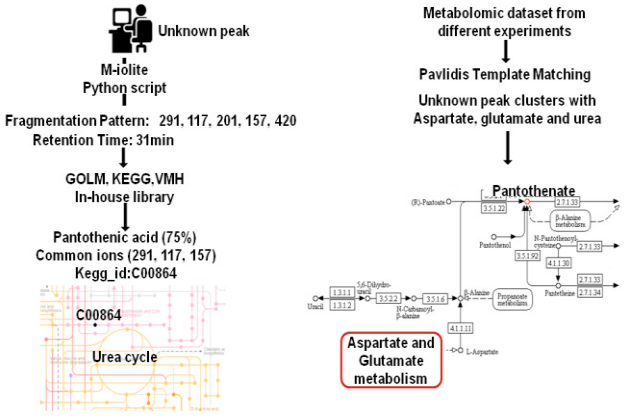

Fig. 2. Unknown Peak identification method based on multiple databases and on the data pattern recognition analysis.

3. CONCLUSIONS

M-IOLITE is an integrated suite for streamlining GC-MS metabolomic data analysis. It provides a user-friendly complete processing workflow through data import, processing, specialized normalization and quality control methods. Also, it integrates the metabolic network analysis for data interpretation and unknown peak identification. In addition, it supports a standardized databank for the repository of the metabolomic data. All the parts have been completed and it will be available online soon upon registration at http://miolite.iceht.forth.gr:8080.

ACKNOWLEDGMENTS

We gratefully acknowledge the TREAT-HEART" research project, No 09SYN-21-965, of General Secretariat for Research and Technology GSRT-GR) "Collaboration" Action/Sub-Action II: "Large-Scale Collaborative Projects" funded by Strategic Reference Framework (NSRF) for funding the PhD fellowship of Ms. Maga-Nteve, the BIOSYS the European Social Fund (ESF) and National Resources of Greece under the Operational Programme "Competitiveness & Entrepreneurship" of the National research project, Action KRIPIS, project No MIS-448301 (2013SE01380036), funded by GSRT-GR and the European Regional Development Fund (Sectoral Operational Programme: Competitiveness and Entrepreneurship, NSRF 2007-2013)/European Commission for partially supporting the purchase of consumables and computer equipment required for the advancement of the project and EC FP7 research project STREPSYNTH (n° 613877).

REFERENCES

Kanani, H., Chrysanthopoulos, P.K. & Klapa, M.I., 2008. Standardizing GC-MS metabolomics. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 871(2), pp.191–201.

Kanani, H. & Klapa, M.I., 2007. Data correction strategy for metabolomics analysis using gas chromatography-mass spectrometry. *Metabolic Engineering*, 9(1), pp.39–51.

Pavlidis, P. & Noble, W.S., 2001. Analysis of strain and regional variation in gene expression in mouse brain. *Genome biology*, 2(10), doi:10.1186/1471-2105-8-240.