Bioinformatics, 31(23), 2015, 3721–3724
doi: 10.1093/bioinformatics/btv475
Advance Access Publication Date: 13 August 2015
Discovery Note

Data and text mining

Determining conserved metabolic biomarkers from a million database queries

Michael E. Kurczy[1], Julijana Ivanisevic[1], Caroline H. Johnson[1], Winnie Uritboonthai[1], Linh Hoang[1], Mingliang Fang[1], Matthew Hicks[1], Anthony Aldebot[1], Duane Rinehart[1], Lisa J. Mellander[2], Ralf Tautenhahn[1], Gary J. Patti[3,4], Mary E. Spilker[5], H. Paul Benton[1] and Gary Siuzdak[1,6,*]

[1]Scripps Center for Metabolomics, The Scripps Research Institute, La Jolla, CA 92037, USA, [2]Department of Physics, University of California San Diego, La Jolla, CA 92093, USA, [3]Department of Chemistry, Washington University in St. Louis, St. Louis, MO 63130, USA, [4]Departments of Genetics and Medicine, Washington University School of Medicine, St. Louis, MO 63130, USA, [5]Pfizer Worldwide Research and Development, San Diego, CA 92121, USA and [6]Departments of Chemistry, Molecular and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on June 18, 2015; revised on July 27, 2015; accepted on August 9, 2015

Abstract

Motivation: Metabolite databases provide a unique window into metabolome research allowing the most commonly searched biomarkers to be catalogued. Omic scale metabolite profiling, or metabolomics, is finding increased utility in biomarker discovery largely driven by improvements in analytical technologies and the concurrent developments in bioinformatics. However, the successful translation of biomarkers into clinical or biologically relevant indicators is limited.

Results: With the aim of improving the discovery of translatable metabolite biomarkers, we present search analytics for over one million METLIN metabolite database queries. The most common metabolites found in METLIN were cross-correlated against XCMS Online, the widely used cloud-based data processing and pathway analysis platform. Analysis of the METLIN and XCMS common metabolite data has two primary implications: these metabolites, might indicate a conserved metabolic response to stressors and, this data may be used to gauge the relative uniqueness of potential biomarkers.

Availability and implementation. METLIN can be accessed by logging on to: https://metlin.scripps.edu

Contact: siuzdak@scripps.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Search analytics is now regularly utilized as an unbiased and anonymous method to survey populations. In a sense each query is a report of some circumstance or set of circumstances that has lead the user to carry out a search. For instance the user might be searching a symptom they are experiencing or the search may reflect a personal bias, in either case this information can be informative within the context of a community. Recent examples include the use of search data to model influenza outbreaks (Ginsberg *et al.*, 2009), and in another case researchers were able to design a rough measure

of regional racism across the United States (Chae *et al.*, 2015). With this in mind and starting from the premise that the majority of users searching the METLIN database are from the population engaged in metabolomic studies we set out to use analytics to survey the broad range of metabolomics research reflected in METLIN search data and reveal the most commonly encountered biomarkers.

The potential impact of biomarker research is far reaching, involving diagnosis, prognosis, drug efficacy and the development of personalized medicine. One particularly intriguing aspect of biomarker discovery is the prospect of reducing the overall cost of health care by increasing the level of medical care through early diagnosis, or through the development of patient specific therapies (Davis *et al.*, 2009). It is therefore understandable that a significant effort has been dedicated to the discovery of unique disease-specific molecules as evidenced by the thousands of biomarker-related papers (Drucker and Krapfenbauer, 2013; Poste, 2011).

The importance of biomarker discovery coupled to the accessibility of new, high sensitivity analytical technology has contributed to this broad interest (Poste, 2011), however a lack of standard best practices has meant that discoveries often have very little practical value. The thousands of papers that have been published on biomarkers have yielded relatively little success (Poste, 2011), for example while the number of biomarker publications have increased by over 20% each year, the number of patent applications for clinical biomarkers remains level (Drucker and Krapfenbauer, 2013). Metabolomics-based biomarker discovery publications are progressively contributing to this body of work and also show the same trends, with the exception of the robust inborn errors of metabolism screens, there are relatively few new clinical metabolite based tests (Xia *et al.*, 2013).

Metabolomic experiments, whether they are accomplished through the use of traditional GC/MS technologies (first demonstrated in the early 1970s by Linus Pauling and Horning; Horning and Horning, 1971; Pauling *et al.*, 1971), or current liquid chromatography mass spectrometry (LC/MS) instrumentation, are well suited for gaining a comprehensive and quantitative view of the metabolome. These approaches have the ability to detect thousands of metabolites from biological samples with high resolution, high sensitivity and a dynamic range typically exceeding four orders of magnitude (Patti *et al.*, 2012). As a result, it is possible to quickly generate datasets rich in metabolite information.

A typical workflow for metabolomics involves several steps; first, mass spectral data is collected for each sample in each sample group. Next the data is analyzed by any number of data processing platforms including Metabolic profiler (Bruker), Simca-P (Umetrics), Markerlynx (Waters), Mass Profiler Pro (Agilent), MetAlign (Lommen, 2009), MZmine (Pluskal *et al.*, 2010), MAVEN (Melamud *et al.*, 2010), MetaboAnalyst (Xia *et al.*, 2012) and XCMS Online (Tautenhahn *et al.*, 2012) to identify the features that significantly change between sample groups. The significantly dysregulated features are then given a putative identification by searching metabolite databases such as METLIN (Supplementary Fig. S1).

METLIN (Smith *et al.*, 2005) is a highly accessed database used for metabolite identification. Since its inception in 2004 METLIN has grown to contain over 240 000 metabolites, 13 000 of which also have MS/MS data in both positive and negative ionization modes at four different collision energies (Supplementary Fig. S1). More than 10 000 users and over 600 citations of the original publication are indicators of its wide application. In addition, METLIN has been integrated with XCMS Online, providing metabolite identifications for the users of this cloud-based data processing platform.

The consistent increase in METLIN use (Supplementary Fig. S1) is largely due to the wide spread adoption of mass spectrometry as an exploratory tool by biologists and biochemists. This has been facilitated by the increased access to mass spectrometry technologies that are sensitive and easy to use. Furthermore, the development of user-friendly bioinformatic platforms and integrated statistical tools (mentioned earlier) has removed a significant technical barrier (Gowda *et al.*, 2014).

Technology has also leveled the playing field so that researchers are typically using similar tools, which for the most part are composed of liquid chromatography coupled to an atmospheric pressure ionization source and either time-of-flight (TOF), quadrupole time-of-flight (QTOF) or quadrupole orbi-trap (Q-Exactive) mass analyzers with a consistent degree of quantitative and mass accuracy. Thus, the masses searched in METLIN provide a representation of the discriminatory metabolites from metabolomic experiments. Here, in order to gain a perspective on the metabolites being searched, METLIN archival data was searched and examined to determine commonly observed metabolites across one million queries.

2 Results and discussion

Researchers using metabolomics are currently investigating a wide array of biological problems, thus the masses searched in METLIN should be a reflection of the diversity of these experiments. However, because metabolism is to some extent conserved we also expect to find commonly dysregulated metabolites to be searched at a high frequency. This assumption was evaluated using Figure 1, which provides an overview of database queries that were submitted by approximately 5000 researchers. This data represents an

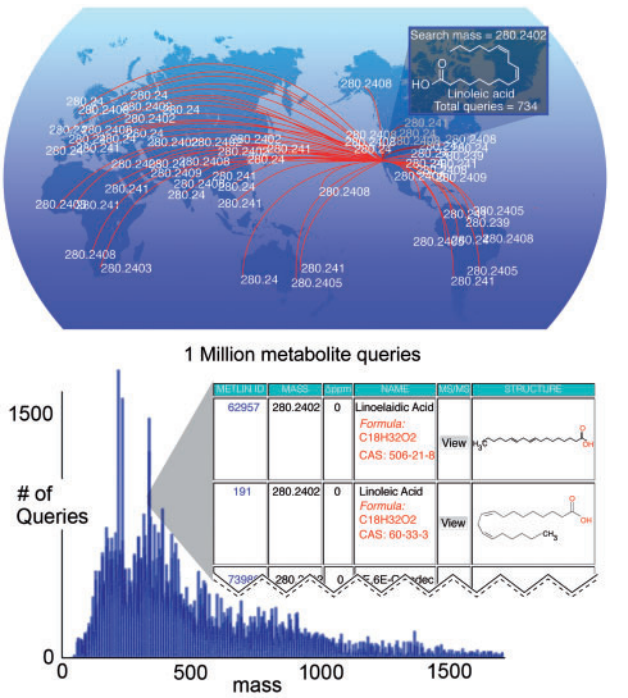

Fig. 1. A cumulative mass spectrum created from a million searches of the METLIN metabolite database. (Top) representation of queries logged from across the globe. (Bottom) The mass spectrum presents the total output of the searches that return METLIN identifications; queries that returned no METLIN IDs have been excluded from this plot (≈9 M). The x-axis represents the *m/z* searched and the y-axis the number of times that particular *m/z* has been searched. The inset shows the output of a METLIN query

aggregate of one million searches that returned METLIN identifications. Interestingly, while a great diversity of metabolites was searched, we found that there are in fact preferentially searched masses. Additionally, over 80% of the searched metabolites fell into a bimodal distribution between the mass ranges of 150 and 450 (Fig. 1), consistent with small molecules that are measured in metabolomics. Because commonly dysregulated metabolite masses will be detected at high frequency and therefore will be searched more often, we suggest that the distribution of this 'mass spectrum' is a manifestation of a conserved metabolic response.

The results obtained from our analysis of METLIN searches produced a list of masses, the number of times each mass has been searched and the number of metabolite IDs that are returned for each query as represented by their neutral mass (Fig. 2). The most frequently searched mass in METLIN was 180.06. We have putatively identified this metabolite as glucose by accurate mass, although this query returns a total of 31 possible metabolites when using the database's default mass accuracy of 30 ppm. This example illustrates an important consideration that is implicit in this kind of meta-data analysis. Some commonly searched masses might be the result of the dysregulation of many isomers as opposed to the dysregulation of one common molecule.

Metabolite identification via MS/MS is still obligatory in metabolomics if one wishes to classify a metabolite as either a specific or a nonspecific signal. Indeed, molecule identification is the most critical aspect of untargeted metabolomics when attempting to understand metabolism. The identification process is outlined in Supplementary Figure S2 where a search returns possible metabolites based on the input mass accuracy and collision energy and metabolite identification is facilitated by the comparison of research MS/MS data to MS/MS data compiled in the METLIN database by matching the precursor mass followed by MS/MS matching based on x-rank (Mylonas *et al.*, 2009). METLIN MS/MS searches were not as frequently performed as accurate mass searches, nor were these searches stored, we instead mined XCMS Online to facilitate the metabolite identifications.

To identify core metabolites and further validate (Supplementary Fig. S2) our assumption that the masses most often searched in METLIN are related to the most commonly dysregulated biomarkers, we also searched meta-data generated by XCMS Online. We were able to cross-compare the METLIN hit list to the most common biomarkers observed in a subset of 2000 XCMS Online experiments. XCMS Online (xcmsonline.scripps.edu) is a web-based platform that was designed to simplify the data analysis of untargeted metabolomic experiments. In 2012 an email was sent out to all XCMS Online users with an invitation to participate in this communal effort to identify shared metabolites across a heterogeneous population of biological specimens and a wide spectrum of phenotypic conditions. Participation in this meta-analysis was done on an opt-in basis only and all data was masked, anonymized and aggregated for statistical analysis. These experiments show that some of the searched masses are most likely due to the appearance of isomers for example while glucose searched at a high frequency it was not commonly dysregulated in real samples.

Using the XCMS data as a guide, we next analyzed a battery of sample types using LC-MS/MS in an attempt to identify the most common biomarkers from real samples. MS/MS data for the most common metabolites encountered in the communal XCMS Online experiments are presented in the Supplementary Material. From the 100 most frequently occurring metabolites with available MS/MS data 24 compounds were identified and listed in Supplementary Table S1, another 24 metabolites were characterized as

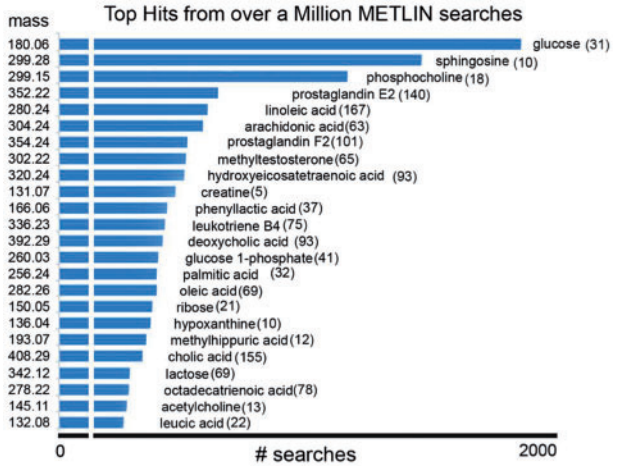

Fig. 2. Top 24 searched metabolites. Putative identifications are based on accurate neutral mass with the number of potential matches are listed in parenthesis

phospholipids. MS/MS spectra were available for an additional 52 compounds but these have not been identified yet (all spectra and tables are available as Supplementary Material). Comparative analysis across XCMS results and the most common METLIN searches revealed a significant overlap. We found that all XCMS results, with exception of 4 metabolites, are within the 90th percentile of the most searched masses in METLIN and half of those in the 90th percentile actually fall within the 99th percentile.

The highest ranked identified molecule in Supplementary Table S1 is linoleic acid. This fatty acid is a useful marker for fat intake and metabolism in humans (Arab, 2003). Although, it is interesting that this molecule would feature so prominently in our list of common metabolites as we can reasonably exclude the possibility that a disproportionate number of XCMS Online users are investigating nutrition. To understand how often this metabolite is reported as a biomarker we performed an ISI web of knowledge search for the name of the metabolite paired with the word *biomarkers*. The search produced 273 separate publications and, following a more detailed survey of the literature, we found that there are at least 10 different conditions linked to the dysregulation of linoleic acid.

The published reports coupled to the MS/MS data shed light on the high occurrence of searches for *m/z* 281.24. We find that linoleic acid is a commonly dysregulated metabolite and while there is little doubt that dysregulation of linoleic acid is discriminatory for all the reported conditions, we would argue that this property makes linoleic acid a relatively non-descript biomarker.

In contrast, METLIN searches corresponding to the biomarkers choline, betaine and trimethylamine N-oxide amount to a total of 25 searches, and *N*,*N*-dimethylsphingosine amounts to 2 search results. The former biomarkers link diet, and gut microflora to cardiovascular disease (Wang *et al.*, 2011) and it might be the number of complicated natural relationships that keep these molecules off of the list of the most common dysregulated metabolites. The latter biomarker, *N*,*N*-dimethylsphingosine, is linked to pain (Patti *et al.*, 2012) but was only found after it was honed in on using a multigroup analysis strategy. In this case the experimental design allowed for the extraction of a subtle metabolite change from the background of nonspecific responses.

Another observation from these experiments is the number of unknowns being observed. The XCMS data are summarized in Figure 3. In this histogram we have plotted the 300 most commonly

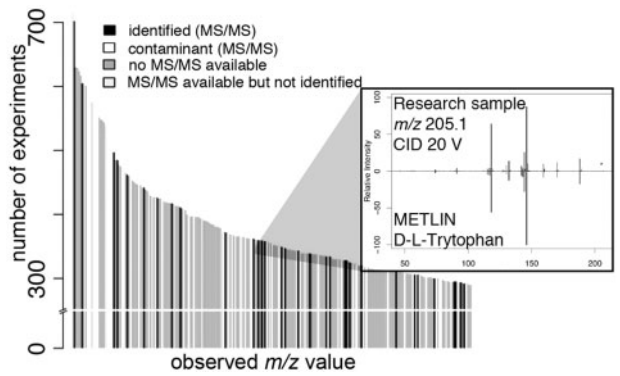

Fig. 3. Observed *m/z* values and frequency for the 300 most common metabolites in the meta-analysis of 2000 untargeted metabolomic experiments across a heterogeneous population of biological specimens and a wide spectrum of phenotypic conditions. The inset shows MS/MS matching the feature corrisponding to tryptophan

dysregulated *m/z* values (the white bars indicate contaminant ions the black bars show identified metabolites the light gray bars are unknowns with MS/MS data and the dark gray bars are unknowns without MS/MS data). The picture given by this plot is that a significant number of the metabolite features that are commonly encountered are unknown. Additionally, we find a large proportion of the histogram represents known contaminates. This finding underscores the importance of proper sample handling and the identification of contamination sources (Weber *et al.*, 2012).

These data are highly dynamic and will evolve with technological advances and changes in standard procedures. As technology will define metabolome coverage, it is likely that the current standard practices will be reflected in the METLIN search data. For example a relatively recent occurrence is the use of hydrophilic interaction chromatography (HILIC) separation technologies coupled to mass spectrometry (Buszewski and Noga, 2012; Spagou *et al.*, 2010). HILIC is used for the analysis of polar metabolites as an alternative to the dominant reversed-phase liquid chromatography (RPLC). Indeed HILIC has been utilized to study central metabolism (Bajad *et al.*, 2006; Heiden *et al.*, 2010; Ivanisevic *et al.*, 2013). Although RPLC is still by far the most utilized separation method, therefore METLIN search data results are likewise primarily driven by RPLC-MS, which generally involve hydrophobic molecules.

3 Concluding remarks

Overall this METLIN-XCMS data provides an opportunity to examine which pathways are commonly related to stress responses and those that reflect more specific responses. For example linoleic acid had a high search frequency in METLIN and appears in 273 manuscripts in the context of biomarkers, a result that is corroborated in our XCMS experimental results making it a conserved biomarker. That is to say it is a general marker for a system under stress that is conserved across many species under various perturbations. By making this data available we are hopefully providing other researchers with a template for biomarker discovery and translation.

Funding

This work was supported by The US National Institutes of Health R01 GM114368 & R01CA170737 and The US Department of Energy DE-AC02-05EH11231

Conflict of Interest: none declared.

References

Arab,L. (2003) Biomarkers of fat and fatty acid intake. *J. Nutr.*, **133**, 925S–932S.

Bajad,S.U. *et al.* (2006) Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *J. Chromatogr. A*, **1125**, 76–88.

Buszewski,B. and Noga,S. (2012) Hydrophilic interaction liquid chromatography (HILIC)-a powerful separation technique. *Anal. Bioanal. Chem.*, **402**, 231–247.

Chae,D.H. *et al.* (2015) Association between an Internet-based measure of area racism and black mortality. *PLoS ONE*, **10**, 4.

Davis,J.C. *et al.* (2009) OUTLOOK The microeconomics of personalized medicine: today's challenge and tomorrow's promise. *Nat. Rev. Drug Discov.*, **8**, 279–286.

Drucker,E. and Krapfenbauer,K. (2013) Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. *EPMA J.*, **4**, 7.

Ginsberg,J. *et al.* (2009) Detecting influenza epidemics using search engine query data. *Nature*, **457**, U1012–U1014.

Gowda,H. *et al.* (2014) Interactive XCMS online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal. Chem.*, **86**, 6931–6939.

Heiden,M.G.V. *et al.* (2010) Evidence for an alternative glycolytic pathway in rapidly proliferating cells. *Science*, **329**, 1492–1499.

Horning,E.C. and Horning,M.G. (1971) Metabolic profiles—gas-phase methods for analysis of metabolites. *Clin. Chem.*, **17**, 802–809.

Ivanisevic,J. *et al.* (2013) Toward omic scale metabolite profiling: a dual separation-mass spectrometry approach for coverage of lipid and central carbon metabolism. *Anal. Chem.*, **85**, 6876–6884.

Lommen,A. (2009) MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.*, **81**, 3079–3086.

Melamud,E. *et al.* (2010) Metabolomic analysis and visualization engine for LC-MS data. *Anal. Chem.*, **82**, 9818–9826.

Mylonas,R. *et al.* (2009) X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry. *Anal. Chem.*, **81**, 7604–7610.

Patti,G.J. *et al.* (2012) Metabolomics implicates altered sphingolipids in chronic pain of neuropathic origin. *Nat. Chem. Biol.*, **8**, 232–234.

Patti,G.J. *et al.* (2012) Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell. Biol.*, **13**, 263–269.

Pauling,L. *et al.* (1971) Quantitative analysis of urine vapor and breath by gas–liquid partition chromatography. *Proc. Natl Acad. Sci. USA*, **68**, 2374–2376.

Pluskal,T. *et al.* (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, **11**, 395.

Poste,G. (2011) Bring on the biomarkers. *Nature*, **469**, 156–157.

Smith,C.A. *et al.* (2005) METLIN: a metabolite mass spectral database. *Ther. Drug Monitoring*, **27**, 747–751.

Spagou,K. *et al.* (2010) Hydrophilic interaction chromatography coupled to MS for metabonomic/metabolomic studies. *J. Sep. Sci.*, **33**, 716–727.

Tautenhahn,R. *et al.* (2012) XCMS online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.*, **84**, 5035–5039.

Wang,Z.N. *et al.* (2011) Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*, **472**, U57–U82.

Weber,R.J.M. *et al.* (2012) MaConDa: a publicly accessible mass spectrometry contaminants database. *Bioinformatics*, **28**, 2856–2857.

Xia,J. *et al.* (2012) MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Res.*, **40**, W127–W133.

Xia,J.G. *et al.* (2013) Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*, **9**, 280–299.