PROCEEDINGS OF THE 9TH INTERNATIONAL CONGRESS OF THERAPEUTIC DRUG MONITORING & CLINICAL TOXICOLOGY, LOUISVILLE, KENTUCKY, APRIL 23–28, 2005

METLIN

A Metabolite Mass Spectral Database

Colin A. Smith, Grace O'Maille, Elizabeth J. Want, Chuan Qin, Sunia A. Trauger, Theodore R. Brandon, Darlene E. Custodio, Ruben Abagyan, and Gary Siuzdak

Abstract: Endogenous metabolites have gained increasing interest over the past 5 years largely for their implications in diagnostic and pharmaceutical biomarker discovery. METLIN (http://metlin.scripps.edu), a freely accessible web-based data repository, has been developed to assist in a broad array of metabolite research and to facilitate metabolite identification through mass analysis. METLIN includes an annotated list of known metabolite structural information that is easily cross-correlated with its catalogue of high-resolution Fourier transform mass spectrometry (FTMS) spectra, tandem mass spectrometry (MS/MS) spectra, and LC/MS data.

Key Words: METLIN, database, metabolites, LC/MS, FTMS, MS/MS

(*Ther Drug Monit* 2005;27:747–751)

Endogenous metabolites represent an important class of biomolecules whose role in fundamental biochemical processes and xenobiotic interactions and potential as biomarkers are in the early stages of being understood.[1–7] There are currently 3 main challenges in metabolite analysis: analytical technology development, chromatographic data processing, and metabolite characterization. The first challenge has largely been met with mass spectrometry, where techniques such as liquid chromatography–mass spectrometry (LC/MS) offer wide dynamic range, high sensitivity, quantitative analysis, and the ability to provide structural information.[8] The second challenge involves processing the significant amount of metabolite mass spectral data generated from biofluid analysis and is a focus of numerous commercial and academic groups. The third challenge has prompted the development of METLIN (METabolite LINk), a metabolite database that incorporates MS data from multiple sources, including high-accuracy FTMS mass measurements and tandem MS data (Fig. 1). With its extensive catalocgue and its multipurpose search capabilities, METLIN stands as a convenient and comprehensive package of valuable resources for characterizing known and unknown metabolites.

The precedent for biologic databases has been established with GenBank (identified genes), Protein Data Bank (3D protein structures), and the Stanford Microarray Database, as they have become an integral part of biologic research.[9–11] METLIN follows the successful model of these widely used databases by incorporating similar features such as free public access, web-based interface, standardized data formats, searchable records, and frequent updates.

A number of mass spectral and metabolite data repositories have previously been created. One such example is the NIST database, a heavily used resource library for electron ionization (EI) mass spectrometry data on over 100,000 compounds (http://www.nist.gov/srd/nist1.htm).[12] However, whereas electrospray ionization (ESI) - based approaches can be used to detect both volatile and nonvolatile compounds, EI is primarily applied to the analysis of volatile compounds. Therefore, the EI component of NIST is relevant for only a fraction of metabolites. Additionally, unlike METLIN, the NIST database is not freely available, and its restricted access is a limiting factor in providing information to researchers not significantly invested in mass spectrometry. Although several NMR "metabonomic" databases also exist, these databases mainly provide collections of combined chemical shift data, and their information on individual molecules is limited.[1,2,13–17]

METLIN OVERVIEW

METLIN is a public, web-based database designed for the archiving, visualization, and analysis of metabolite data. Its objective is to provide the following information from multiple biologic sources.

1. Structural and physical data on known endogenous metabolites, drugs, and drug metabolites (100–1200 daltons)
2. High-accuracy FTMS data from reference biofluid/tissue samples
3. Reference tandem MS data from known metabolites and metabolite derivatives
4. LC/MS profiles from primarily human and some model organisms

METLIN allows instant retrieval of FTMS, MS/MS, and LC/MS mass analysis results by enabling its user to directly query the database via user-specified parameters. For instance, when a search is performed on plasma LC/MS data for ions of mass-to-charge ratio (*m/z*) 456, a chromatogram of the matching ion is returned along with any known metabolites of equal mass. As an additional resource for the user in identifying or characterizing an observed ion, METLIN provides high-accuracy mass measurement data that are readily available for

Received for publication June 1, 2005; accepted June 23, 2005.
From The Scripps Research Institute, Molecular Biology and Center for Mass Spectrometry, La Jolla, California 92037, USA.
Reprints: Gary Siuzdak, The Scripps Research Institute, Molecular Biology and Center for Mass Spectrometry, La Jolla, CA 92037 (e-mail: siuzdak@scripps.edu).

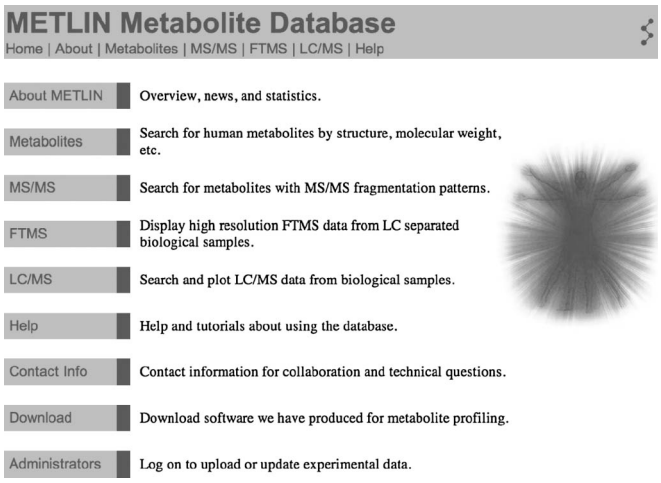


FIGURE 1. METLIN's homepage showing multiple links to metabolite searches, such as metabolite structure, MS/MS fragmentation, FTMS, and LC/MS data.

retrieval. Search results can be visualized with plotting options that include scatter plots, histograms, and 3D analysis. In addition, color coding of mass values based on a number of different parameters is possible. The METLIN database is updated frequently and can be accessed at http://metlin.scripps.edu/.

Known Endogenous and Drug Metabolites

The METLIN database maintains a growing catalogue of known metabolites. Each metabolite is annotated with its chemical formula and structure to supplement and aid users' understanding of its MS and (when available) ion fragmentation data. Researchers can also identify structurally similar metabolites using METLIN's chemical substructure searching capability (Fig. 2). In addition, query results in METLIN provide each metabolite with its CAS tag number and a direct link to its entry in the Kyoto Encyclopedia of Genes and Genomes (KEGG).

Although the expansion of METLIN's metabolite resources is ongoing, it already possesses the content, framework, and functionality of a powerful metabolite information database. METLIN's ability to facilitate the identification and characterization of metabolite ions based on a combination of both physical properties (eg, mass value, chemical formula) and sample specificity (eg, biofluids from healthy and disease states) offers a unique opportunity to the endogenous metabolite research community to benefit tremendously from its remarkable potential.

FTMS and MS/MS Data

One of the major challenges in metabolite research is the characterization of potential biomarkers. METLIN assists researchers in meeting this challenge by providing public access to high-resolution FTMS data, MS/MS data, and a growing compilation of known metabolites.

FTMS analysis is an important aspect of metabolite characterization. The high resolution (>50,000) offered by FTMS provides the possibility of generating high-accuracy *m/z* measurements (less than 1 ppm). FTMS measurements are frequently used to hypothesize elemental compositions of observed mass values. When compared with theoretical isotope distributions (Fig. 3), precise isotopic patterns obtained from FTMS analysis can further confirm elemental formulas of unknown ions. On METLIN, a user can query FTMS data using a chemical formula, and METLIN will automatically display the relevant *m/z* range and superimpose a theoretical isotopic pattern for the entered formula. METLIN currently houses FTMS spectra from a number of chromatographically separated human serum fractions. The number of FTMS spectra is continuously expanding as additional FTMS data sets from other biofluids and tissues are being routinely incorporated.

In addition to its compilation of metabolite FTMS data, METLIN now offers searchable tandem MS data (when available) that can be retrieved with simple user-defined queries. The MS/MS data consist of the metabolite MS/MS pattern as well as the intensity and resolution of each MS fragment (Fig. 4). By specifying a precursor mass range on the MS/MS search form, a researcher can easily reference the MS/MS profile of known metabolites in METLIN against the MS/MS profile of an unknown compound. The tandem MS data METLIN provides are therefore particularly valuable to researchers who are in the early stages of metabolite identification.

LC/MS Data

METLIN currently catalogues more than 200 biofluid and tissue analyses that are available for qualitative analysis. To support a diverse array of cross-correlation studies, METLIN tracks multiple characteristics associated with each sample including variables such as patient/animal type, species, age,

52 metabolite(s) found

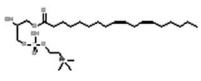
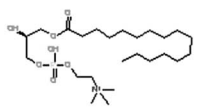
Mass	Name	Formula	CAS	KEGG	Structure
520.3398	1-Linoleoylglycerophosphocholine	$C_{26}H_{51}NO_7P$	15895-41-7	C04100	
496.3398	1-Palmitoylglycerophosphocholine	$C_{24}H_{51}NO_7P$	17659-62-0	C04102	

FIGURE 2. Result of a search for metabolites containing a phosphate. Standard output includes calculated exact mass, name, formula, CAS registry number, KEGG compound ID (if applicable), and metabolite structure.

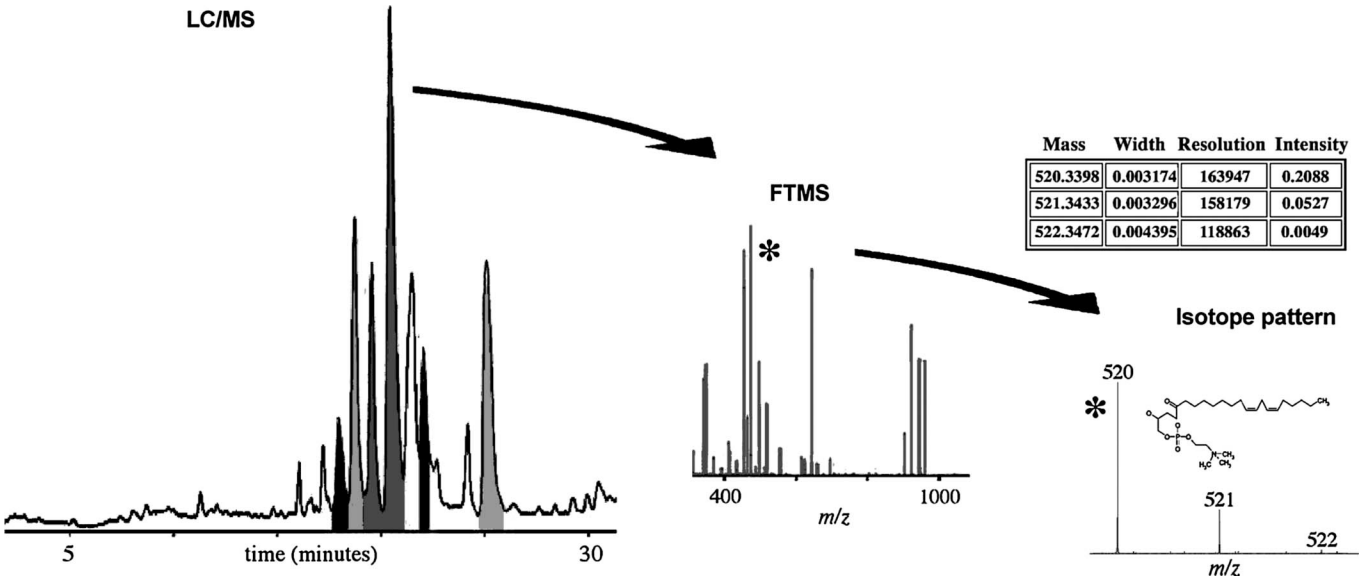


FIGURE 3. LC/MS data from a human plasma sample including FTMS data on a fraction (the second colored peak) and an enlargement of a particular mass range showing the isotope pattern of 1-linoleoylglycerophosphocholine, a high-abundance metabolite found in plasma membranes. The high-resolution mass spectrum isotopic pattern for $C_{26}H_{51}NO_7P^+$ is shown. The FTMS spectrum comes from fraction 2 of the human serum sample chromatogram. This figure represents a composite of actual images and text from the web site.

gender, ethnicity, disease, disease severity, and genotype. Other sample variables include biofluid/tissue source, preparation, and metabolite derivatization method. Analytical variables such as chromatography, mass spectrometry instrument type, and ionization mode are also included in METLIN's metabolite profile.

In the design of METLIN, a data management strategy was specifically developed for handling a large amount of data, which can range in size from 10 to 200 megabytes per data file depending on the LC gradient length, mass accuracy, and sample rate.[18] Such data require extensive processing and reduction, without which it would be difficult to analyze.[19,20] Even in its reduced form, efficient data evaluation and searching demand a standard format for data storage and access.[21] The implementation of METLIN avoids the computationally intensive task of handling large datasets by storing only significant peaks from the mass spectra. Significant peaks are those that remain after robust noise filtering and relative thresholding. This important reduction step decreases the data storage and processing requirements of the database by more than 1000-fold. Currently METLIN employs a previously described algorithm[19] for LC/MS peak finding. Each peak is represented as a 3-dimensional data point consisting of its retention value (ie, % acetonitrile), *m/z*, and intensity. Most data sets with reverse-phase LC/MS using normal flow rates (~250 μL/min) for electrospray ionization yield approximately 400–1200 peaks, averaging about 25 kilobytes in size.

Recognition that peaks from different samples with similar coordinates correspond to the same mass value is a central challenge in metabolite profiling.[18] Further, noise and retention time shifts make accurate automated clustering a challenge. This somewhat parallels problems in gene and protein expression experiments, where cross-hybridization and peptide fragment identification are significant issues,[22,23] but is exacerbated in LC/MS because many of the observed metabolites have not yet been identified. METLIN begins to address that problem by clustering peaks into a presumptive list of ions, which can then be placed into more straightforward comparisons.

Equipped with a robust set of search features, METLIN provides easy means for users to locate data of interest among an extensive catalogue of metabolite information and MS data. Queries on METLIN are simple and straightforward. For instance, a user can easily search all LC/MS data from a given type of cancer with a single mouse click. For more advanced searches, users can opt for multiple search parameters, such as specific *m/z* and retention time ranges, ionization mode, sample properties, or biofluid type. Researchers can search and view every sample stored in the database, displaying information such as age, gender, disease state, biofluid or tissue type, ionization mode, and the number of peaks.

Implementation

The METLIN metabolite database is implemented using the open-source software tools, MySQL (http://www.mysql.com/) and PHP (http://www.php.net/). MySQL is a relational database system ideally suited for very fast retrieval of relatively static records. METLIN is currently stored in 4 linked tables. The chemical substructure and similarity search makes use of the MolCart MySQL plug-in (http://www.molsoft.com). The PHP scripting language was designed from the outset for web programming and has been extensively used for data-driven web sites. The MySQL/PHP combination has

(Metabolites 1-3 of 3) Change Query

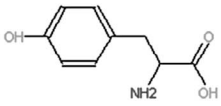
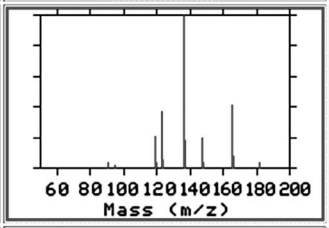
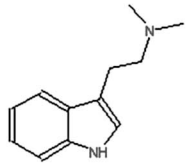
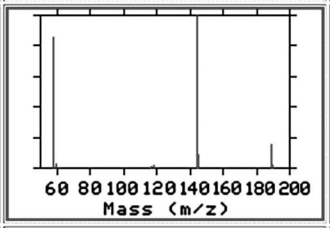
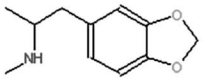
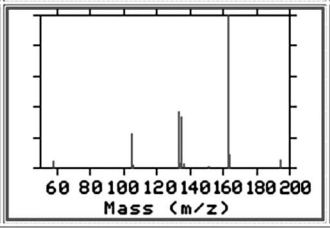
MID	Mass	Name	Formula	Structure	MS/MS Plot	MS/MS Fragments
34	181.0739	Tyrosine	$C_9H_{11}NO_3$			13
69	188.1313	N,N-Dimethyltryptamine (DMT)	$C_{12}H_{16}N_2$			8
3311	193.1103	3,4-Methylenedioxymethamphetamine (MDMA)	$C_{11}H_{15}NO_2$			11

FIGURE 4. Result of a search for MS/MS profiles of metabolites with mass range 180–200 daltons. Standard output includes calculated exact mass, name, formula, structure, MS/MS profile, and number of MS fragments.

previously been leveraged for a scalable and robust DNA microarray database.[24]

Because of its high resolution, FTMS *m/z* and intensity data are stored in binary, flat file format. A custom engine was written to quickly parse those files for graphic output and peak identification. All interactive graphics are produced on the fly with the GNUPLOT graphics package (http://www.gnuplot.info/).

DISCUSSION AND CONCLUSION

Current high-throughput techniques in gene and protein expression can capture a great deal of information about the molecular state of cells and biofluids. However, these methods can not predict the actual metabolic consequences of upstream changes. Posttranslational events, allosteric regulation, and other factors can have profound effects that are not easily measured with DNA microarray and proteomic assays. Advances in small-molecule metabolite profiling will help improve our understanding of these biologic processes and the molecular basis of drugs and disease.

Development of gene and protein expression assays has depended extensively on well-developed genome and proteome databases. METLIN, a metabolite database based on mass spectral data, achieves a similar role in increasing our ability to understand the metabolome. METLIN currently addresses the needs of metabolite and biomarker researchers with 3 main components: (1) a database of known endogenous metabolites, drugs, and drug metabolites and their respective structures, (2) high-resolution FTMS data and MS/MS data from a number of standard biologic samples, and (3) a database of LC/MS profiles from a broad spectrum of samples and processing methods. With these tools and data in place, the primary goal of METLIN is to facilitate metabolite identification by providing known information on metabolites as well as data from multiple biologic sources including urine, plasma, tears, and cerebrospinal fluid. The combination of data storage, peak assignment, and identification will expedite the identification of key metabolites in healthy and diseased states and will greatly enhance the area of metabolite research.

ACKNOWLEDGMENTS

The authors appreciate support from NIH grants 5P30 EY012598-04 and 5R24EY01474-04, we also thank MolSoft LLC for providing the MolCart chemical cartridge.

REFERENCES

1. Gelpi E. Biomedical and biochemical applications of liquid chromatography–mass spectrometry. *J Chromatogr A*. 1995;703:59–80.
2. Volk KJ, Hill SE, Kerns EH, Lee MS. Profiling degradants of paclitaxel using liquid chromatography–mass spectrometry and liquid chromatography–tandem mass spectrometry substructural techniques. *J Chromatogr B Biomed Sci Appl*. 1997;696:99–115.
3. Chace DH. Mass spectrometry in the clinical laboratory. *Chem Rev*. 2001; 101:445–477.
4. Buchholz A, Hurlebaus J, Wandrey C, Takors R. Metabolomics: quantification of intracellular metabolite dynamics. *Biomol Eng*. 2002;19:5–15.

5. Watkins SM, German JB. Toward the implementation of metabolomic assessments of human health and nutrition. *Curr Opin Biotechnol*. 2002; 13:512–516.
6. German JB, Roberts MA, Watkins SM. Personal metabolomics as a next generation nutritional assessment. *J Nutr.* 2003;133:4260–4266.
7. Forrester JS, Milne SB, Ivanova PT, Brown HA. Computational lipidomics: a multiplexed analysis of dynamic changes in membrane lipid composition during signal transduction. *Mol Pharmacol*. 2004;65:813–821.
8. Anari MR, Sanchez RI, Bakhtiar R, et al. Integration of knowledge-based metabolic predictions with liquid chromatography data-dependent tandem mass spectrometry for drug metabolism studies: application to studies on the biotransformation of indinavir. *Anal Chem*. 2004;76:823–832.
9. Sherlock G, Hernandez-Boussard T, Kasarskis A, et al. The Stanford Microarray Database. *Nucleic Acids Res*. 2001;29:152–155.
10. Galperin MY. "The Molecular Biology Database Collection: 2004 update." Database issue. *Nucleic Acids Res*. 2004;32:D3–22.
11. Wheeler DL, Church DM, Edgar R, et al. "Database resources of the National Center for Biotechnology Information: update." Database issue. *Nucleic Acids Res*. 2004;32:D35–D40.
12. Wagner C, Sefkow M, Kopka J. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry.* 2003;62:887–900.
13. Holmes E, Shockcor JP. Accelerated toxicity screening using NMR and pattern recognition-based methods. *Curr Opin Drug Discov Devel*. 2000; 3:72–78.
14. Holmes E, Nicholson JK, Tranter G. Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. *Chem Res Toxicol*. 2001;14:182–191.
15. Shockcor JP, Holmes E. Metabonomic applications in toxicity screening and disease diagnosis. *Curr Top Med Chem*. 2002;2:35–51.
16. Beckonert O, Bollard ME, Ebbels TMD, et al. NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. *Anal Chim Acta*. 2003;490:3–15.
17. Ebbels T, Keun H, Beckonert O, et al. Toxicity classification from metabonomic data using a density superposition approach: "CLOUDS." *Anal Chim Acta*. 2003;490:109–122.
18. Frenzel T, Miller A, Enzel KH. A methodology for automated comparative analysis of metabolite profiling data. *Eur Food Res Technol*. 2003;216:335–342.
19. Hastings CA, Norton SM, Roy S. New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. *Rapid Commun Mass Spectrom*. 2002;16:462–467.
20. Andreev VP, Rejtar T, Chen HS, et al. A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. *Anal Chem*. 2003;75:6314–6326.
21. Mendes P. Emerging bioinformatics for the metabolome. *Brief Bioinform*. 2002;3:134–145.
22. Clauser KR, Baker P, Burlingame AL Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem*. 1999;71:2871–2882.
23. Lee I, Dombkowski AA, Athey BD. Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray. *Nucleic Acids Res*. 2004;32:681–690.
24. Saal LH, Troein C, Vallon-Christerssen J. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol*. 2002;3(8):SOFTWARE0003.

