A New Technique (COMSPARI) to Facilitate the Identification of Minor Compounds in Complex Mixtures by GC/MS and LC/MS: Tools for the Visualization of Matched Datasets

Jonathan E. Katz, Darren S. Dumlao, and Steven Clarke
Department of Chemistry and Biochemistry and the Molecular Biology Institute, Los Angeles, California, USA

Jörg Hau
Nestlé Research Center, Nestec Ltd., Lausanne, Switzerland

In the rapidly growing field of metabolomics, it is common to analyze complex biological samples by chromatography coupled to mass spectrometry. While several techniques are available for the detection of significant peaks in individual samples, it is still difficult to determine small differences between similar samples. Using conventional software, visual inspections of individual chromatograms or individual mass spectra are often of little use because the differences in the composition of small molecules are too small to be recognizable. Thus, we developed a new approach to visualizing mass spectral datasets using a tool that allows one to easily detect these small differences between mass spectra and chromatograms derived from matched samples. Using these tools on extracts from wild-type and methyltransferase knockout strains of the yeast *Saccharomyces cerevisiae*, we were able to readily identify those mass spectra in our data sets that were different between the wild-type and the knockout extracts and to identify the molecules involved. The software was also successfully applied to a set of LC/MS data from peptide digests that were performed with identical substrates but different enzymes. We have named this visualization tool COMSPARI (COMparision of SPectrAl Retention Information) and are making the software publicly available via Internet at http://www.biomechanic.org/comspari/. (J Am Soc Mass Spectrom 2004, 15, 580–584) © 2004 American Society for Mass Spectrometry

In recent years, it has become common to analyze complex samples by chromatography coupled to mass spectrometry. In many of these cases, in particular with samples from biological systems, emphasis is not on the identification of all individual compounds but on the *difference* between two datasets from two different samples. For example, it is often desirable to elucidate the differences in the molecular composition of cells under various conditions or that have distinct genetic backgrounds.

Several problems are inherent to this task. First of all, GC and in particular LC separations that are coupled to detection by MS often result in complex datasets with a substantial level of noise. Using conventional background subtraction techniques or "differential chromatograms" is of limited success, as the small, statistically occurring differences in ion abundance or in retention time may lead to erroneous peaks.

A second problem lies in the size of many datasets; as an example, a typical LC/MS run to characterize a peptide digest covers a mass range of about 2000 *m/z* units and has about 3000 individual scans. This makes the manual inspection of all the individual spectra or chromatograms with current methodologies very time-consuming and tedious.

We faced these problems when we studied TMT1, the *Saccharomyces cerevisiae* homologue of the *Escherichia coli trans*-aconitate methyltransferase [1]. Like *E. coli*, *S. cerevisiae* was shown to methylate *trans*-aconitate endogenously but there was also a second, more abundant, substrate as determined by HPLC fractionation of radioactively labeled cell extracts [2]. To characterize this unidentified alternate substrate, cell extracts of parental and TMT1 knockout strains of *S. cerevisiae* were prepared and analyzed by GC/MS. However, using conventional software, we found no obvious differences between the wild-type and knockout chromatograms.

Published online February 10, 2004

Address reprint requests to Dr. S. Clarke, Molecular Biology Institute, UCLA, 611 Charles E. Young Drive East, Los Angeles, CA 90095-1570, USA. E-mail: clarke@mbi.ucla.edu

1044-0305/04/$30.00
doi:10.1016/j.jasms.2003.12.011

Received August 25, 2003
Revised December 18, 2003
Accepted December 18, 2003

To investigate whether a difference in the spectra was present but obscured, several techniques were tested.

Early tools to aid in the interpretation of such noisy datasets include the use of a two-dimensional plot of the complete run as a "map" or "eagle's view" [3]. This display is very suitable for the analysis of an individual sample at a time, but if two such "maps" are to be overlaid, the multitude of information therein will make it difficult to compare the two datasets in detail, in particular if one is searching for minor differences in huge datasets.

An automatic and dynamic signal selection procedure was published in 1996 by Windig and coworkers [4]. Their Component Detection Algorithm (CODA) selects mass chromatograms with both low noise and low background by calculating a "similarity index" between each raw mass chromatogram and its smoothed and mean-subtracted version. Used with LC/MS datasets, CODA can reduce the number of mass chromatograms to be investigated roughly by one order of magnitude without significant loss of information but is usually tailored to the investigation of single datasets.

Another approach, particularly suited for GC/MS, is to sort the GC/MS files in 14-ion series [5], which reduces a large dataset down to only 14 mass traces. This technique is frequently used to identify compounds responsible for a contamination or other differences in flavor chemistry, but is less useful in LC/MS due to the high background observed with that technique which will easily obscure small signals.

The AMDIS software [6, 7] is also primarily used to process GC/MS data. AMDIS was developed to take chromatograms of complex samples, identify all the separate components that are present and then, using a deconvolution algorithm, extract the EI spectra for each of the present components. Again, this approach aims at single-sample datasets, and the peak detection algorithm behind AMDIS requires the presence of several mass traces that converge at the same retention time. This latter condition is rarely fulfilled with LC/MS runs, as the main information is often concentrated into one or a few *m/z* values.

Thus, although the techniques described above are useful for the investigation of individual runs, they are less practical for the comparison of multiple datasets. Indeed all the software that was available to us, either from the instrument manufacturer or from other sources, proved to be insufficient for facile inspection of "paired" data from our matched samples.

Therefore, we have developed COMSPARI (COMparision of SPectrAl Retention Information), a new approach to visualize mass spectral datasets. It can present the data from a pair of GC/MS or LC/MS runs in an easily accessible and informative mirrored display. The use of a similar plot to show differences between datasets has long been used for the comparison of spectra to reference compounds (e.g., [8, 9]) and for chromatogram comparisons [5]. However, the application presented here effectively makes use of the dynamic variation of such a display when browsing through a dataset. We also developed a complementary preprocessing utility, *cdf2ascii*, to convert the original data files from the common NetCDF format into a more easily processed format.

We show here the usefulness of this software in identifying the yeast endogenous methyltransferase substrate from GC/MS data. Additionally, we show the general utility of the method by analyzing LC/MS data from peptide digests done with identical substrates but different enzymes.

This software is made freely and publicly available under the terms and conditions of the GNU Public License [10] and can be downloaded via Internet at http://www.biomechanic.org/comspari/.

Methods

GC/MS

The tmt1::KanMX4 *S. cerevisiae* strain, HCY001, and the TMT1 overexpression strain, JK5, were previously described [2]. Small molecule cell extracts were prepared as described in [2] and derivatized with *N,O*-bis(trimethylsilyl)trifluoroacetamide supplemented with 10% trimethylchlorosilane (BSTFA + 10% TMCS, Pierce Biotechnology, Rockford, IL). One μL samples were separated using an Agilent (Palo Alto, CA) 6890+ gas chromatograph equipped with a HP-5MS column (Agilent) attached to a Micromass (Manchester, UK) GCT time of flight mass spectrometer with an electron impact (EI) source. Mass spectra were acquired in positive ion mode over the scan range *m/z* 50 to 800 at a rate of 2 scans per second. Data acquisition and data evaluation were performed using the Micromass MassLynx software, version 3.5.

LC/MS

The peptide samples used here were derived from a proprietary peptide, which was digested under identical conditions but using two different enzymes. Analyses were performed on an UltiMate HPLC system (LC Packings, Amsterdam, The Netherlands) coupled to a Micromass QToF-2 mass spectrometer equipped with an electrospray ion source. One μL of the samples was injected onto a C-18 MB250/1 100-5 HPLC column (Macherey-Nagel, Oensingen, Switzerland). Solvent A consisted of water with 1% formic acid, Solvent B was acetonitrile with 1% formic acid. A flow rate of 50 μl/min was used, and gradient elution was performed from 2% B at 10 min to 100% B at 50 min. Mass spectra were acquired in profile mode by scanning over the mass range *m/z* 150 to 2000 with a scan time of 1 s. Data acquisition and data evaluation were performed using MassLynx v4.0.

A full description of the GC/MS and LC/MS meth-

odologies can be found at http://www.biomechanic.org/comspari/.

File Format Conversions

The original data were converted from the native Micromass file format to netCDF by the Micromass supplied "databridge" tool. While netCDF is a standard format for data exchange, it is not as suitable for rapid processing: the netCDF data files are essentially a sequence of mass spectra, and thus reconstitutions of mass chromatograms requires one to read *all* the mass spectra sequentially and process them one by one. Therefore, we found it necessary to write a preprocessor to convert data files from netCDF into a "simpler" format. This preprocessor, *cdf2ascii* ("CDF to ASCII"), takes a netCDF file and writes out all mass spectra and mass chromatograms as plain ASCII files. These can then easily be processed further, not only by COMSPARI but also by other standard command-line utilities such as *grep*, *awk*, *sort*, etc. The converter *cdf2ascii* is distributed with the COMSPARI package and is based on Jörg Hau's *CDFread* software [11] and David Stranz' original public ANDI-MS netCDF implementation [12].

COMSPARI

The data visualization software COMSPARI has a simple, prompt-based interface and uses the *gnuplot* plotting package [13] for display. Both *cdf2ascii* and COMSPARI can be compiled and used on any recent computing platform, in particular Linux as well as all current versions of the Microsoft Windows operating system.

Results

COMSPARI Allows for Easy Data File Visualization

COMSPARI has two primary modes of operation: "mass spectrum" and "selected ion chromatogram". Upon program launch, COMSPARI expects the names of two (optionally only one) datasets to be given, and then displays a set of paired mass spectra in a head-to-tail plot. We found that once an initial matched pair of mass spectra is displayed, the most informative searching technique is to switch to the selected ion chromatogram mode (for example, at ion 50, "c 50") and walk through the displays (simply by pressing the "Enter" key) while occasionally adjusting the number of *m/z* traces that are displayed simultaneously (e.g., "w 5"). The user can then perform a number of display operations, such as changing the displayed pair of mass traces, "zooming" into any part of the display, using different magnification factors for the two axes, annotating peaks and printing to PostScript files. Most of these operations are performed intuitively either by entering data with the keyboard, or by using the mouse. Regions where there are marked differences in the *m/z* chromatograms can be further explored by switching to the corresponding mass spectrum at the scan of interest (e.g., "s 308"), and switching between the "spectrum" and "chromatogram" view is done with a simple keystroke command.

A particular feature of COMSPARI is the capability of visualizing either a single data file or two data files in a head to tail plot. When visualizing two data files, either plot can be intensity scaled (command "z" or "Z") or shifted (offset, "o" or "O") to adjust for run-to-run variation.

Spectral Differences are Readily Visualized by COMSPARI

After conversion of the GC/MS data files from the two "matched" samples (TMT1 wild-type and knockout extracts), the datasets were visualized with the COMSPARI software. Figure 1a shows a view from COMSPARI in chromatogram mode, showing the high similarity between the samples for two arbitrarily selected *m/z* values. Figure 1b, however, shows a clear peak (at *m/z* 319) that is only present in the parent strain extract (scan 308).

Figure 1c shows the mass spectrum at scan 308, demonstrating a set of distinct peaks in the wild-type spectrum and not present in the knockout spectrum. Thus, "walking" through the mass chromatograms in selected ion mode clearly reveals the presence of compounds that are present in only one of the two matched samples. Although the total ion chromatogram and most of the selected ion chromatograms are visually almost identical (Figure 1a), a novel compound can be identified by one or more selected ion traces (Figure 1b). The process is truly intuitive, and the "mirrored" display allows one to distinguish differences visually even if there is a small shift, for example, in the intensity of the surrounding peaks or in the retention time.

The region of difference illustrated in Figure 1c was then deconvoluted using the AMDIS package [6, 7] to separate the EI spectrum of our unknown from the EI spectra of overlapping compounds, and the resultant spectrum was used to ultimately identify the unknown methyltransferase product as the methyl ester of 3-isopropylmalate.

Another field of application is the data evaluation of samples analyzed by LC/MS. One of the authors faced the problem of analyzing two similar samples with minor differences by LC/MS. The samples were derived from a proprietary peptide that was digested using two different enzymes but under otherwise identical conditions, and the task was to assess the differences between the two peptides. CODA processing [3] yielded more than 300 mass traces for each sample where significant peaks would be present. Albeit this represents a reduction by a factor of six compared to the

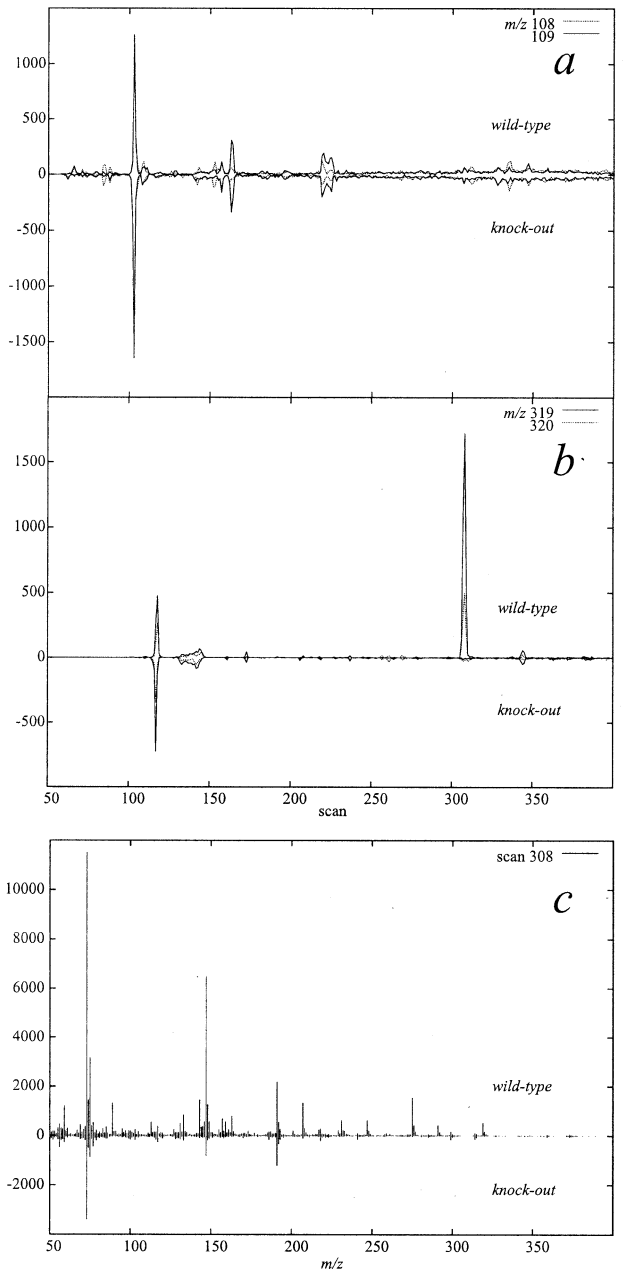

Figure 1. The use of COMSPARI. Matched parent and TMT1 knockout strain yeast extracts were fractionated, BSTFA derivatized, subjected to GC/MS analysis and visualized with COMSPARI. Panel (**a**) shows a view from COMSPARI in selected ion chromatogram mode, width 2, showing the high similarity between the samples for a random sampling of selected ions (in this case 108 and 109). Panel (**b**) shows a very clear peak (*m/z* 319) that is present in the parent strain derived extract at approximately scan 308 and is *not* present in the knockout strain derived extracts. Panel (**c**) is a spectrum display from COMSPARI at scan 308 showing that there is a set of very distinct peaks in the wild-type not present in the knockout.

raw dataset, it would still be a challenging task to identify differences by manual inspection of all the individual mass traces. In contrast to this, COMSPARI allows a quick comparison of these samples, as shown in Figure 2. The trace of *m/z* 309 exhibits some peaks

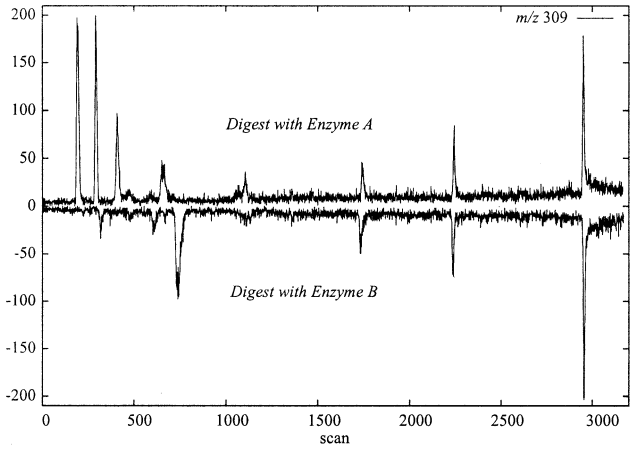

Figure 2. Comparison of differential digests of the sample peptide. The sample peptide was digested with two enzymes ("A" and "B") under otherwise identical conditions and subjected to LC/MS as described in the methods. Shown is the data visualized with COMSPARI; the trace of *m/z* 309 exhibits some peaks that are present in both digests (around scans 1730, 2240, 2950), but also several peaks that are formed exclusively with one of the two enzymes (scans 190, 290, and 650 for Enzyme A; scan 730 for Enzyme B).

that are present in both digests (around scans 1730, 2240, 2950), but also several peaks that are formed exclusively with one of the two enzymes (scans 190, 290 and 650 for Enzyme A, scan 730 for Enzyme B). If the investigation is limited to the mass traces that are preselected by CODA processing as described above, the two datasets can be compared within relatively short time and the corresponding peaks investigated, for example, by MS/MS or accurate mass measurements.

Discussion

COMSPARI was developed to compare "matched samples" for possible differences in intensity that occur simultaneously in the time and *m/z* domains. We have found that our tool facilitates identifying even small differences in chromatograms of such paired samples. The mirrored display helps to detect differences, as two similar mass spectra—or mass chromatograms—with similar peak positions and peak abundances will exhibit a symmetrical pattern, with the baseline situated roughly in the center of the display. Most of the time the display will thus be balanced, but if one of the two samples displayed has a distinct peak, either the baseline will be offset or at least the overall visual appearance will be perturbed. Thus, the operator can track down differences quickly and intuitively.

In our study, the "classical" manual inspection of EI GC/MS spectra of wild-type compared to knockout extracts turned out to be uninformative and lengthy. However, manual inspection using COMSPARI yielded the difference within only 15 minutes. Similar improve-

ments in processing efficiency are observed in our LC/MS peptide analysis example.

Thus, COMSPARI is a valuable tool for the comparison of "matched" datasets. Its use should be beneficial particularly in the field of metabolomics, where the aim is often to determine the difference between two closely related yet different samples. By making our tool publicly available (http://www.biomechanic.org/comspari/, both source and binary distributions are available), we hope that other researchers will be able to benefit from this package.

Acknowledgments

This research was supported by grant GM26020 from the National Institutes of Health and by shared equipment grant CHE-007829: (GCT) from the National Science Foundation. The authors thank Dr. Kym Faull and Dr. Julian Whitelegge from the Pasarow Mass Spectrometry Laboratory at UCLA for their support and expertise. They also thank Lionel Bovetto and François Masson, Nestlé Research Center Lausanne, for the preparation of the peptide digests.

References

1. Cai, H.; Clarke, S. A Novel Methyltransferase Catalyzes the Methyl Esterification of *trans*-Aconitate in *Escherichia coli*. *J. Biol. Chem.* **1999,** *274,* 13470–13479.
2. Cai, H.; Dumlao, D.; Katz, J. E.; Clarke, S. Identification of the Gene and Characterization of the Activity of the *trans*-Aconitate Methyltransferase from *Saccharomyces cerevisiae*. *Biochemistry* **2001,** *40,* 13699–13709.
3. Hau, J.; Linscheid, M. MSGraph: A Program for the Display of LC/MS Data. *Spectrochim. Acta* **1993,** *48B*, E1047-E1051. Updated version with CODA processing at http://homepage.sunrise.ch/mysunrise.ch/joerg.hau/sci/, accessed December 2003.
4. Windig, W.; Phalp, J. M.; Payne, A. W. A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry. *Anal. Biochem.* **1996,** *68,* 3602–3606.
5. Fay, L. B.; Staempfli, A. A. New Approach to Processing of Gas Chromatographic/Mass Spectrometric Data for Detection of Off Flavors in Complex Mixtures. *J. AOAC Int.* **1995,** *78,* 1429–1434.
6. Stein, S. An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data. *J. Am. Soc. Mass Spectrom.* **1999,** *10,* 770–781.
7. AMDIS website, http://chemdata.nist.gov/mass-spc/amdis/ accessed December 2003.
8. Ando, T.; Nyhan, W. L.; Bachmann, C.; Rasmussen, K.; Scott, R.; Smith, E. K. Isovaleric Acidemia: Identification of Isovalerate, Isovalerylgycine, and 3-Hydroxylisovalerate in Urine of a Patient Previously Reported as Having Butyric and Hexanoic Acidemia. *J. Pediatrics* **1973,** *82,* 243–248.
9. NIST MS Search Program (software version 2.0) website, http://www.nist.gov/srd/nist1a.htm accessed December 2003.
10. GNU General Public License website, http://www.gnu.org/copyleft/gpl.html accessed December 2003.
11. CDFread website, http://homepage.sunrise.ch/mysunrise.ch/joerg.hau/sci/ accessed December 2003.
12. ANDI-MS (D. Stranz) FTP site, ftp://ftp.sjo.appliedbiosystems.com/ accessed August 2003.
13. Project:gnuplotdevelop:summary website, http://sourceforge.net/projects/gnuplot/ accessed December 2003.