Systematic Identification of Conserved Metabolites in GC/MS Data for Metabolomics and Biomarker Discovery

ARTICLE *in* ANALYTICAL CHEMISTRY · MARCH 2007

Impact Factor: 5.64 · DOI: 10.1021/ac0614846 · Source: PubMed

Mark Styczynski

Georgia Institute of Technology

Joel Moxley

Massachusetts Institute of Technology

Anal. Chem. **2007,** *79,* 966−973

Systematic Identification of Conserved Metabolites in GC/MS Data for Metabolomics and Biomarker Discovery

Mark P. Styczynski, Joel F. Moxley, Lily V. Tong, Jason L. Walther, Kyle L. Jensen, and Gregory N. Stephanopoulos*

Department of Chemical Engineering, Massachusetts Institute of Technology, Room 56-469c, Cambridge, Massachusetts 02139

Analysis of metabolomic profiling data from gas chromatography−mass spectrometry (GC/MS) measurements usually relies upon reference libraries of metabolite mass spectra to structurally identify and track metabolites. In general, techniques to enumerate and track unidentified metabolites are nonsystematic and require manual curation. We present a method and software implementation, freely available at http://spectconnect.mit.edu, that can systematically detect components that are conserved across samples without the need for a reference library or manual curation. We validate this approach by correctly identifying the components in a known mixture and the discriminating components in a spiked mixture. Finally, we demonstrate an application of this approach with a brief analysis of the *Escherichia coli* metabolome. By systematically cataloguing conserved metabolite peaks prior to data analysis methods, our approach broadens the scope of metabolomics and facilitates biomarker discovery.

The goal of metabolomics—the metabolite analog of genomics and proteomics—is the measurement of concentrations (or "metabolite profiles") of as many cellular metabolites as possible, usually with applications to functional genomics.[1,2] Certain aspects of metabolomics suggest that exhaustive metabolite profiling may be possible: the number of known metabolites present in many organisms (e.g., yeast) is 10- to 100-fold fewer than the number of genes or proteins,[3−5] and the cost of measuring these metabolites is by comparison significantly lower. To date, the coupling of metabolomic data with other cell-wide data has yielded valuable insight into underlying biochemical processes and has contributed to numerous advances in the area of functional genomics.[6−9] However, obstacles to exhaustive metabolite profiling persist, one of the most significant being the chemical diversity of metabolites. Unlike DNA or proteins, metabolites do not adhere to a subunit-based chemistry, so assaying for many metabolites (with many chemistries[10]) simultaneously is difficult. Gas chromatography−mass spectrometry (GC/MS) is one method frequently used to assay for a variety of metabolites, and the aim of this work is to improve the downstream analysis of this GC/MS data independent of upstream experimental protocols.

Common Data Processing Methods. First, current methods for analyzing metabolomic GC/MS data must be understood. In targeted GC/MS analysis, when only the concentrations of a few specific compounds are desired, only certain regions of the chromatogram or certain *m*/*z* values of mass spectra may be considered relevant. For nontargeted metabolomic analyses, on the other hand, the entire chromatogram (for all *m*/*z* values) is important, prompting efforts to select experimental parameters that maximize metabolite peak accessibility.[11] However, accurate analysis of all of this data presents some challenges. Some "real" chromatographic peaks may be hard to distinguish from noise. Other peaks may contain mixtures of metabolites that are coeluting, so the individual MS scans of their peaks are not pure spectra of either component. Thus, the immediate processing steps after the storage of raw GC/MS data typically include peak enumeration (distinguishing "true" peaks from noise in a chromatogram) and, with increasing frequency in recent literature, spectral deconvolution (obtaining putative pure spectra from two overlapping peaks). These steps may be performed either with proprietary software for a specific manufacturer's apparatus or with publicly available software like AMDIS.[12] Whether or not spectral deconvolution is applied, the user is left with a set of chromatographic peaks that putatively represent individual components in the sample mixture. Often, multiple experimental conditions have

* Corresponding author. Fax: (617)-253-3122. E-mail: gregstep@mit.edu.

[1] Sumner, L. W.; Mendes, P.; Dixon, R. A. *Phytochemistry* **2003**, *62*.
[2] Fiehn, O. *Plant Mol. Biol.* **2002**, *48*.
[3] Forster, J.; Famili, I.; Fu, P.; Palsson, B. O.; Nielsen, J. *Genome Res.* **2003**, *13*, 244−253.
[4] Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000**, *28*, 27−30.
[5] Mewes, H. W.; Albermann, K.; Bahr, M.; Frishman, D.; Gleissner, A.; Hani, J.; Heumann, K.; Kleine, K.; Maierl, A.; Oliver, S. G.; Pfeiffer, F.; Zollner, A. *Nature* **1997**, *387*, 7−65.
[6] Raamsdonk, L. M.; Teusink, B.; Broadhurst, D.; Zhang, N.; Hayes, A.; Walsh, M. C.; Berden, J. A.; Brindle, K. M.; Kell, D. B.; Rowland, J. J.; Westerhoff, H. V.; van Dam, K.; Oliver, S. G. *Nat. Biotechnol.* **2001**, *19*, 45−50.
[7] Weckwerth, W.; Loureiro, M. E.; Wenzel, K.; Fiehn, O. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7809−7814.
[8] Hirai, M. Y.; Yano, M.; Goodenowe, D. B.; Kanaya, S.; Kimura, T.; Awazuhara, M.; Arita, M.; Fujiwara, T.; Saito, K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 10205−10210.
[9] Hellerstein, M. K. *Metab. Eng.* **2004**, *6*, 85−100.
[10] Kanani, H.; Klapa, M. *Metab. Eng.*, in press.
[11] O'Hagan, S.; Dunn, W. B.; Brown, M.; Knowles, J. D.; Kell, D. B. *Anal. Chem.* **2005**, *77*, 290−303.
[12] Stein, S. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 770−781.

10.1021/ac0614846 CCC: $37.00
Published on Web 12/24/2006

been assayed, and all of the results need to be analyzed and compared.

It is in the following steps that methods for metabolomic experimental data analysis are rather scattered. Few tools are available for easy comparisons of multiple experimental conditions; ChromaTOF (LECO, St. Joseph, MI) is a representative example of some of the better proprietary software for performing this analysis. However, ChromaTOF can only be used with LECO mass spectrometers, so such a program is not useful for many scientists. Similar problems arise with other manufacturers' software packages; many are inadequate for whole-data-set analysis, and analysts typically do not have the luxury of buying new equipment strictly for its superior software. Alternatively, there are other proprietary programs available that work for a variety of manufacturers' hardware, such as MetAlign (Plant Research International, Wageningen, The Netherlands). However, such programs are typically expensive and (as yet) have not seen wide use in scientific literature. There are a few freely available packages that can perform some of this data analysis, such as mzMine[13] and SpecAlign,[14] although these newer tools also have yet to see widespread testing in scientific literature. Rather, a much more common approach is to compare the spectra from a given run to a reference library of spectra so that metabolite peaks may be tracked by names.

In metabolomic experiments using this approach, the scope of most automated data analysis techniques is limited to those compounds which have been isolated, been purified, and had their spectra stored in a reference library. However, even the largest publicly available libraries[15–17] are often incomplete, leaving many metabolite peaks unidentifiable without additional experimental work.[18] To avoid such issues, one may create a customized reference library using standard reagents; however, this is prohibitively labor-intensive and the resulting library will necessarily be incomplete because many compounds of interest are not commercially available (e.g., only 200 of the 600 known yeast metabolites can be purchased[19]). Alternatively, supplementing a pre-existing library by adding every spectrum from every experimental run is an equally undesirable option: subsequent matches to this "complete" library have less value, as spectra generated by noise and other biologically irrelevant factors would overtake the true set of cellular components.

Tracking of Unidentified Metabolite Peaks. When a chromatographic peak's spectrum cannot be matched to any reference library entry, the analyst must then decide the disposition of this unidentifiable spectrum. It is not uncommon for such spectra to be discarded. In cases where some of these spectra are retained, this is usually the result of a labor-intensive, somewhat ad hoc process of comparing multiple chromatograms to a reference chromatogram. Table 1 enumerates recent GC/MS metabolite profiling studies, some of which relied on nonsystematic, manually driven curation of unidentified metabolite peaks. While previously mentioned tools such as ChromaTOF, Metalign, and mzMine are capable of allowing some tracking of unknowns, this process still requires significant intervention on behalf of the user and may not allow for completely consistent, automated tracking of unknown metabolite peaks. Perhaps more importantly, these methods typically entail comparison to one single reference spectrum. However, if a spectrum may contain significant noise (see discussion below and Figure 1), the use of any one given spectrum as a reference for all others may be undesirable, as the "true" low-concentration metabolites may be difficult to parse out from the data. Thus, while unidentifiable peaks are sometimes retained in GC/MS data analysis, an automated, systematic method for including these peaks in subsequent analysis and appropriately accounting for noisy spectra is as desirable a goal for GC/MS metabolomic analysis as it was for NMR analysis.[6–8]

Table 1. Usage and Cataloging of Unidentified GC/MS Peaks in Metabolomic Studies[a]

author	total peaks	identified peaks	unidentified peaks used?	catalog method
Fiehn 2000[35]	326	164	yes	nonsystematic[b]
Roessner 2001[36]	88	61	yes	nonsystematic[c]
Taylor 2002[37]	>400	90	yes	nonsystematic[b]
Roessner-Tunali 2003[38]	?	73	yes	nonsystematic[c]
Verdonk 2003[39]	?	62	no	-
Prithiviraj 2003[40]	?	253	no	-
Duran 2003[41]	?	?	yes	nonsystematic[d]
Morris 2004[42]	60	27	yes	nonsystematic[e]
Barsch 2004[43]	~200	65	yes	nonsystematic[e]
Strelkov 2004[44]	330	164	yes	nonsystematic[b]
Vikram 2004[45]	1081	49	no	-
Bino 2005[46]	~7500	320	yes	nonsystematic[f]
Devantier 2005[47]	?	29	no	-
Kaplan 2005[48]	416	81	yes	nonsystematic[b]
Desbrosses 2005[49]	~500	87	yes	nonsystematic[b]
Tarpley 2005[50]	?	21	no	-
Herebian 2005[51]	>200	130	yes	nonsystematic[b]
Tikunov 2005[52]	322	100	yes	nonsystematic[f]
Villas-Bôas 2005[53]	?	45	no	-
Brosché 2005[54]	?	22	no	-
this work	554	51	yes	systematic

[a] A literature survey reveals that of those recent GC/MS metabolite profiling studies that include unidentified peaks, most rely on nonsystematic, manually assisted curation of those unidentified peaks. [b–e] Peaks are cataloged by applying the AMDIS,[b] Thermoquest-MassLab,[c] MSFACTS,[d] or Xcalibur[e] program to a reference chromatogram. As stated in the text, this class of methods fails to use a systematic conservation criterion and requires manual curation to scale to moderate to large data without becoming dominated by noise. [f] Peaks are analyzed by MetAlign, a method described in our text; while the unknown peak data are systematically analyzed, the true metabolites are not cataloged.

Another goal of metabolomics research is to enumerate metabolites known as biomarkers that discriminate classes of samples obtained from different cellular conditions by being absent, present, or differentially present. As noted above, there are presently few methods for comparing many sets of spectra in order to identify components characteristic of a sample or group of samples. A commonly used method is principal component analysis (PCA);[20–22] however, the results of PCA (the loadings)

(13) Katajamaa, M.; Oresic, M. *BMC Bioinf.* **2005**, *6*, 179.

(14) Wong, J. W. H.; Cagney, G.; Cartwright, H. M. *Bioinformatics* **2005**, *21*.

(15) Fiehn, O. *Phytochemistry* **2003**, *62*, 875−886.

(16) Nielsen, K. F.; Smedsgaard, J. *J. Chromatogr., A* **2003**, *1002*, 111−136.

(17) Ausloos, P.; Clifton, C. L.; Lias, S. G.; Mikaya, A. I.; Stein, S. E.; Tchekhovskoi, D. V.; Sparkman, O. D.; Zaikin, V.; Zhu, D. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 287−299.

(18) Fiehn, O.; Kopka, J.; Trethewey, R. N.; Willmitzer, L. *Anal. Chem.* **2000**, *72*, 3573−3580.

(19) Dunn, W. B.; Bailey, N. J.; Johnson, H. E. *Analyst* **2005**, *130*, 606−625.

(20) Villas-Boas, S. G.; Mas, S.; Akesson, M.; Smedsgaard, J.; Nielsen, J. *Mass Spectrom. Rev.* **2005**, *24*, 613−646.

(21) Duran, A. L.; Yang, J.; Wang, L.; Sumner, L. W. *Bioinformatics* **2003**, *19*, 2283−2293.

(22) Jonsson, P.; Gullberg, J.; Nordstrom, A.; Kusano, M.; Kowalczyk, M.; Sjostrom, M.; Moritz, T. *Anal. Chem.* **2004**, *76*, 1738−1745.

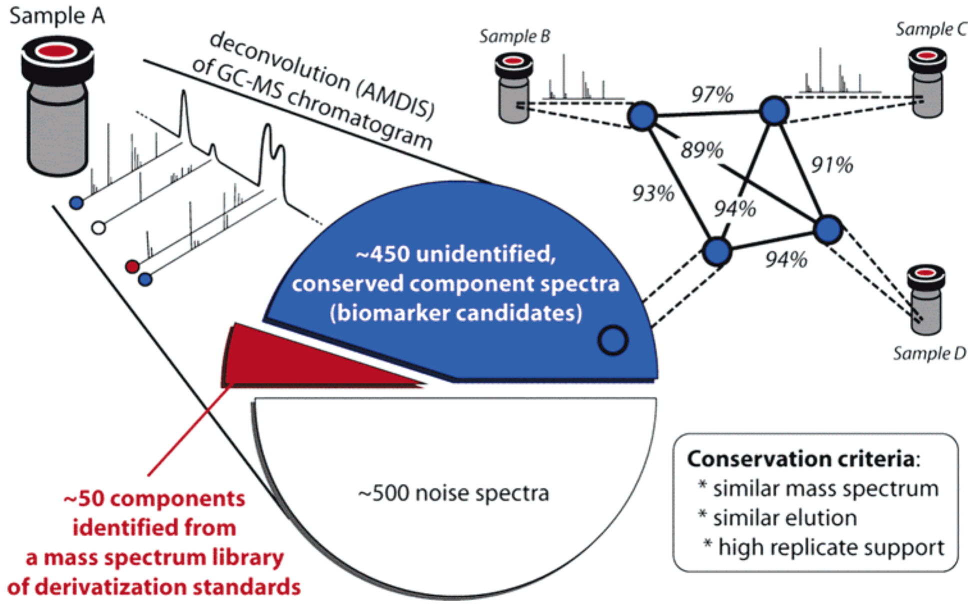
Sample A
deconvolution (AMDIS) of GC-MS chromatogram
Sample B
Sample C
Sample D
97%
89%
91%
93%
94%
94%
~450 unidentified, conserved component spectra (biomarker candidates)
~500 noise spectra
~50 components identified from a mass spectrum library of derivatization standards
Conservation criteria:
* similar mass spectrum
* similar elution
* high replicate support

Figure 1. The composition of a typical GC/MS metabolomic sample motivates the need to systematically track the potential biomarker components not found in a library of standards. For the GC/MS vial labeled "Sample A", we display the deconvolution and enumeration of components by AMDIS from the GC/MS output. The large number of "noise" components (usually small peaks near the GC baseline) emphasizes that constructing a conglomerate library of all putative components would be unwise. Components labeled with blue circles did not match any spectra in a standard library but, as displayed in the upper right, had highly similar spectra occur in replicate samples. This is not surprising, as construction of standard spectra libraries are confounded by a variety of factors. Given the conservation of this blue component, we conclude that it is not noise, and thus may catalog its spectrum in a SpectConnect library. In future analysis, we can track this specific component among conditions and investigate its potential biomarker activity. Note that though the relative amount of "white" noise spectra may decrease with better deconvolution and peak enumeration, there will still remain a significant number of blue "unidentifiable" spectra without an automated, systematic way to handle them.

are often not intuitive, and the optimized function (capturing variance) is not necessarily ideal for biomarker discovery. As previously mentioned, Metalign and other software can perform pairwise spectrum comparison in place of PCA, and other methods also exist[13,23−25] to these ends, although most either focus on a small subset of data or are not generally applicable. The method most relevant to this work[26,27] avoids dependence on libraries of standards, but still depends on complex multivariate analysis; it focuses solely on biomarkers rather than identification and tracking of all possible metabolites in the samples. As such, biomarker discovery remains an open problem, with numerous potential applications in diagnosis and prognosis that reach beyond the scope of a simple classification.[28]

The SpectConnect Approach. Here, we present a method and freely available implementation, SpectConnect, to automatically catalog and track otherwise unidentifiable conserved metabolite peaks across sample replicates and different sample condition groups without use of reference spectra. SpectConnect compares every spectrum in each sample to the spectra in every other sample. By doing so, it is capable of determining which components are conserved (according to some criteria) across replicate samples. SpectConnect also determines which of these components differentiate one sample condition (e.g., time or treatment) from another, whether by changes in concentrations or merely by their presence/absence. The only requirement of the experimental measurements is that each sample condition must have replicates. In a sense, SpectConnect relies on an increase in signal relative to noise that is created by this requirement of replicates. While injection ("technical") replicates are the easiest way to provide the required replicates, it is also desirable to include biological replicates in a group of samples. This is due to systematic error in peak detection and deconvolution software that may consistently find a noise peak in technical replicates. Though this approach adds more biological variability to a group in terms of metabolite concentrations, it should have significantly less impact on the presence/absence of a metabolite. Ultimately, we hypothesize that these "true", important spectra will be conserved across most or all replicates of a sample, while spectra that are artifacts of noise will not.

For the core of the necessary computations, SpectConnect uses Gemoda, a freely available generic motif discovery algorithm for sequential data.[29] Gemoda efficiently compares candidate spectra and identifies conserved spectra across samples using various clustering methods. If a chromatographic peak is true signal and not noise, we expect that the mass spectrum of each of its occurrences in the different samples will be pairwise similar to each other. Projected onto a graph consisting of nodes for each

[23] Nielsen, N. P.; Smedsgaard, J.; Frisvad, J. C. *Anal. Chem.* **1999**, *71*, 727−735.

[24] Katz, J. E.; Dumlao, D. S.; Clarke, S.; Hau, J. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 580−584.

[25] Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C.; Lamers, R. J.; van der Greef, J.; Timmerman, M. E. *Bioinformatics* **2005**, *21*, 3043−3048.

[26] Idborg, H.; Edlund, P. O.; Jacobsson, S. P. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 944−954.

[27] Idborg-Bjorkman, H.; Edlund, P. O.; Kvalheim, O. M.; Schuppe-Koistinen, I.; Jacobsson, S. P. *Anal. Chem.* **2003**, *75*, 4784−4792.

[28] Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. *Nat. Rev. Drug Discovery* **2002**, *1*, 153−161.

[29] Jensen, K. L.; Styczynski, M. P.; Rigoutsos, I.; Stephanopoulos, G. N. *Bioinformatics* **2006**, *22*.

spectrum and edges between each pair of spectra that are similar, this pairwise similarity defines a cluster known as a "clique". By limiting our clusters to cliques, we exclude potential artifacts of noise and focus on what we believe are the true metabolites in the sample. We exploit Gemoda's ability to efficiently find cliques in order to locate these clusters of spectra that represent the true metabolites. An illustration of this process can be seen in Figure 1.

With SpectConnect, we can find three distinct types of information. First, we identify all of the components that are conserved across a single group of samples or replicates. Second, for each metabolite peak conserved in at least one group, we can determine all other groups in which it occurs. Finally, for any given pair of groups, we can find the likelihood, to some degree of statistical significance, that each metabolite peak is present in unequal amounts in the groups.

METHODS

Experimental Methods. Standard mixtures were prepared using commercially available stock chemicals. *Escherichia coli* strains obtained from previous work[30] were fermented and sampled over 30 h. The samples were quenched and frozen; after thawing, the metabolites were separated from the rest of the mixture and concentrated. MCF derivatization[31] was performed prior to GC/MS analysis using an HP 5890 GC coupled to an HP 5971 quadrupole mass selective detector (EI). Additional details on all experimental and statistical protocols used can be found in Supporting Information.

Peak Identification and Deconvolution. AMDIS[12] was used to perform component peak identification and spectral deconvolution. The following parameters were used: medium shape requirement, low sensitivity, and medium resolution. Each GC/MS sample's results were processed, and the ELU files created as output of AMDIS (and containing data for all of the enumerated peaks in each sample) were saved for use with SpectConnect. Though in principle any program could be used for component peak enumeration, we designed this implementation of SpectConnect to work with AMDIS output since AMDIS is freely available and accepts a wide variety of manufacturers' raw data formats. However, AMDIS's tendency toward false positives in an effort to be sensitive[12] makes subsequent clustering steps much more difficult and is one reason that SpectConnect performs "preprocessing" steps, as mentioned below. However, even if peak enumeration and deconvolution were ideal, the resulting data would still greatly benefit from a systematic search for conserved metabolite peaks and biomarker candidates.

SpectConnect. SpectConnect is implemented in Python (http://www.python.org) and uses Gemoda[29] for the majority of its computations. A user may choose how well two spectra must match to be deemed "similar", how much error should be allowed in retention time for occurrences of the same metabolite peak, and how many times a peak must occur to be considered "conserved" (also known as its "support"). This study used the default values for these parameters, which are as follows: spectra must be 80% similar on the basis of a weighted dot product[12] to be considered similar, they must be within 1 min in retention time to be considered similar, and they must occur in at least 75% of sample replicates to be considered conserved. The weighted dot product for our spectrum-to-spectrum comparison is defined as

$$\frac{(\sum_m m^2\sqrt{I_{1,m}I_{2,m}})^2}{(\sum_m m^2 I_{1,m})(\sum_m m^2 I_{2,m})}$$

where m takes on all valid m/z values in either mass spectrum and $I_{1,m}$ and $I_{2,m}$ are the ion intensities at $m/z = m$ for the first and second spectra being compared, respectively. Additional (optional) restrictions can also be placed on the data analysis, including minimum relative abundances for peaks. Further analysis of parameter selection and resulting limitations is given in the Discussion section.

Given a set of ELU files representing replicates from the same sample condition, SpectConnect first parses each file to extract all pertinent information, including the retention time and mass spectrum of each peak. It then uses Gemoda to identify and eliminate all internal matches from each sample: any spectra within a single sample that are within the elution threshold and meet the weighted dot product similarity criterion are combined into a single group, and one single spectrum is chosen as the representative peak for further comparisons and clustering.

After all replicate input files have been parsed and preprocessed, the resulting information is used to find conserved metabolite peaks. Each sample's reduced set of spectra (which now includes unique spectra that are not internally similar) is supplied to Gemoda for clustering. Here, the appropriate elution, similarity, and support thresholds are enforced. As stated earlier, we require that true metabolite peaks have spectra that are well-conserved across replicate samples. We use Gemoda's maximal clique-finding algorithm to find sets of pairwise similar spectra. However, our requirements can lead to complications because clique-based clustering does not require that each item participate in just one cluster. Since a given metabolite peak may be involved in multiple clusters, it is important to minimize overlapping similarities and similarities that are not believed to be as significant. This desire to ensure a library of non-similar metabolite peaks, combined with AMDIS's high-sensitivity and low-specificity approach to peak identification, makes the previously explained preprocessing steps necessary.

For each clique returned by Gemoda, the most representative spectrum is chosen, as judged by a weighted dot product of spectra. Since the cliques may be overlapping, representative spectra may be self-similar, and so this set of spectra is further processed to eliminate internal matches in the same way as described above. After this step, SpectConnect creates a final "library" of metabolite peaks conserved for this sample condition; this library is returned to the user as output and is used in additional calculations.

The identification of conserved peaks is repeated for each "sample condition" or set of injection or biological replicates, resulting in a set of peak libraries. These sample libraries can then be further condensed into one cumulative library by using Gemoda to determine in how many conditions each metabolite

[30] Alper, H.; Miyaoku, K.; Stephanopoulos, G. *Nat. Biotechnol.* **2005**, *23*, 612–616.

[31] Villas-Boas, S. G.; Delicado, D. G.; Akesson, M.; Nielsen, J. *Anal. Biochem.* **2003**, *322*, 134–138.

peak occurs. The resulting library is essentially the union of all previous libraries, and it is also returned to the user as output.

RESULTS

Mixtures of Known Components. To verify SpectConnect's ability to enumerate individual components of a mixture, we analyzed a known standard mixture of amino acids with MCF derivatization in replicate GC/MS runs; the results of this experiment can be found in Supporting Information Table 1. Of the standard mixture components, 16 should have been detectable using MCF derivatization; using SpectConnect, we identified 15 of them. Isoleucine and leucine could not be simultaneosly identified due to the resolution of our preprocessing technique (see Methods section): though they are properly deconvolved by AMDIS, their spectra are too similar to be classified as distinct by SpectConnect. Thirty-two additional peaks were also detected as conserved across replicates, of which some were identified as byproducts of the derivatization reagents. We believe that the remainder of these peaks, which are 10- to 100-fold smaller in size than the median of the known components' peaks, reflect impurities in the stock mixture.

We next analyzed the same mixture spiked with additional compounds. Using the same amino acid standard as above, we spiked eight stock chemical compounds into the standard (see Supporting Information Table 1) and used MCF derivatization to analyze the resulting sample in replicate GC/MS runs. Each of these compounds was identified by SpectConnect as occurring exclusively in the supplemented mixture. In addition, approximately 100 other peaks were identified as conserved in the supplemented mixture and not present in the control mixture. On the basis of analysis of single-compound GC/MS runs and the fact that these peaks are also 10- to 100-fold smaller than the median peak of the known components, we are confident that these extra peaks largely represent impurities introduced with the addition of multiple doping compounds. Some of these peaks were even identified from our library of standards: for instance, in the spiked sample we found conserved peaks for *cis*-aconitate, a dehydration product of citric acid and thus a reasonable "contaminant". It should also be noted that aspartic acid (from the standard mixture) and citric acid (added to the standard mixture) coelute, yet we still identified both as unique mixture components and *only citric acid as discriminatory between sample groups*.

By perturbing each SpectConnect parameter from its default, we evaluated the impact of parameter selections on our results (see Supporting Information Table 2). This impact was fairly small for this example; the primary impact was on the number of other, unexpected conserved peaks that were found. All parameter choices but one were able to detect the conserved and differential metabolites in the two samples.

Biological Samples. To demonstrate the capabilities of SpectConnect on biological samples, we analyzed GC/MS time-course data from fermentation runs conducted with three different strains of *E. coli* (see Supporting Information Methods section). The strains came from previous work in engineering the overproduction of lycopene.[30] We found that across five time points over 30 h, there were a total of 544 metabolite peaks (chromatogram peaks) that occurred in at least one of the strains in at least one time point (but not in a blank derivatization control), while 184 of those occurred in all of the strains in at least one time point. (Parameter sensitivity for the number of conserved peaks found is addressed in the Discussion section.) Qualitatively, this indicates that the genetic differences of these strains have caused significant differences in their respective metabolisms. This result is to be expected for mutants with deletions of metabolic enzymes: some subsets of metabolites are rendered inaccessible, so a significant metabolic adjustment is necessary to compensate for such changes.

Using this cumulative library of 544 metabolite peaks, we then analyzed the metabolomic profile of one mutant strain relative to that of the reference strain in the course of the fed-batch cultivation. Figure 2 shows, for the time sample with the greatest metabolic deviation from the reference strain as measured by lycopene production (strain 1 at 24 h),[30] the increase in the number of metabolite peaks that are detected as biomarkers when using the SpectConnect library relative to those detected when using a pre-existing library of reference spectra for MCF-derivatized metabolites. The *p*-values plotted are the results of *t*-tests comparing relative abundances (abundances normalized by total ion count) of each metabolite peak between the two sample conditions. (These tests are performed automatically by SpectConnect to identify biomarker candidates.) Only molecules that can be detected in at least one replicate of each sample are included in Figure 2. While a few compounds are identified using the known library, significantly more spectral signatures are detected with SpectConnect.

Figure 3 demonstrates that principal components analysis using the SpectConnect library for the data qualitatively captures the differences between strains and time points better than the previously known MCF reference library. The SpectConnect library allows better resolution between the two strains and even suggests more distinct "trajectories" along sample times than is possible using the MCF reference library, including a divergence of trajectories before a difference in lycopene production is detectable. Since PCA is unsupervised, these results support the notion that the metabolite peaks returned by SpectConnect capture biological aspects of the system rather than just noise. As noted above, PCA is not the ideal method for identifying potential biomarkers; rather, Figure 2 indicates our putative biomarkers as defined by pairwise *t*-tests. Figure 3 serves to support the notion that the additional metabolite peaks detected by SpectConnect help to better capture the variance in the data.

DISCUSSION

The method and program that we present here allows for systematic, automated analysis of GC/MS metabolite profiling data sets, including metabolites that may not be structurally identified by a reference library. In the *E. coli* metabolome data set, we see almost an order of magnitude increase in the number of metabolite peaks that can be tracked with GC/MS measurements without manual curation of unidentified peaks. Accordingly, the enriched metabolite peak set allowed for a more fruitful downstream data analysis: differential relative abundances identified more biomarker candidates than would be possible using strictly library-based approaches without unidentified peaks, and PCA projections offered better characterization and separation of different sample classes. In addition, we note that all SpectConnect computations

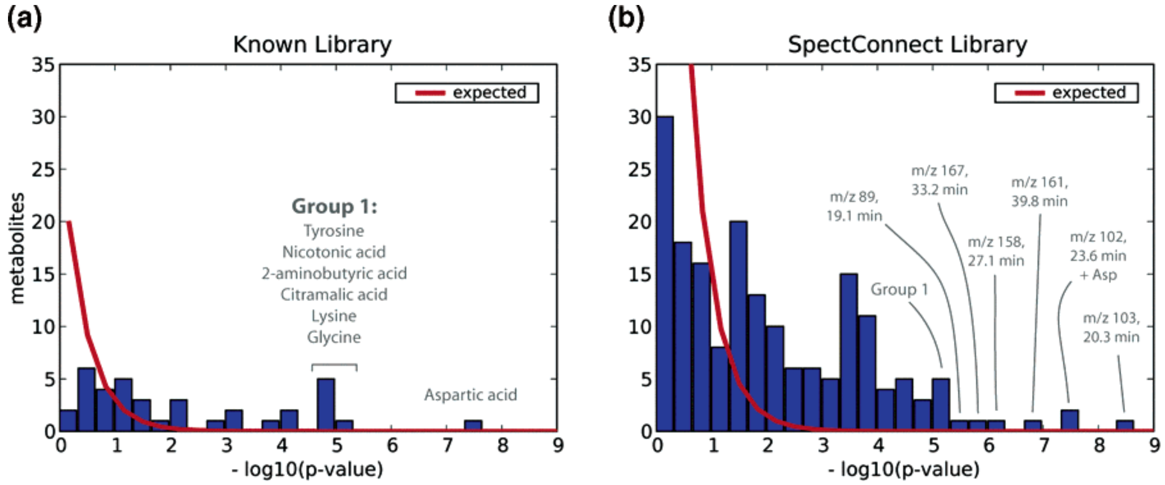

Figure 2. Comparing conditions using a global SpectConnect library helps identify more biomarker candidates. In each graph, we compare wild-type and mutant (strain 1) cell populations sampled from their respective fed-batch reactors at the point of largest deviation of the phenotype of interest (lycopene production at fermentation time 24 h). In each graph we use the identical set of experimental measurements to calculate confidences in differential levels (*p*-value) among the samples for those metabolite peaks which were detected in both conditions. (We do not display data for those peaks which may be considered biomarkers merely by their presence or absence.) Lines illustrating the expected number of metabolite peaks in each bin are based on a uniform random distribution of *p*-values. (A) Using a reference library for MCF-derivatized metabolites, previous methods would only be able to identify the biomarker candidates in this *p*-value histogram. (B) After using SpectConnect to create a larger library of conserved components and their spectra, we can identify a greater number of statistically significant biomarkers, including all of the most significant biomarkers found using the known library. Interesting biomarker candidates may then be further explored and biologically identified using known ad hoc techniques involving the spectra and retention information.

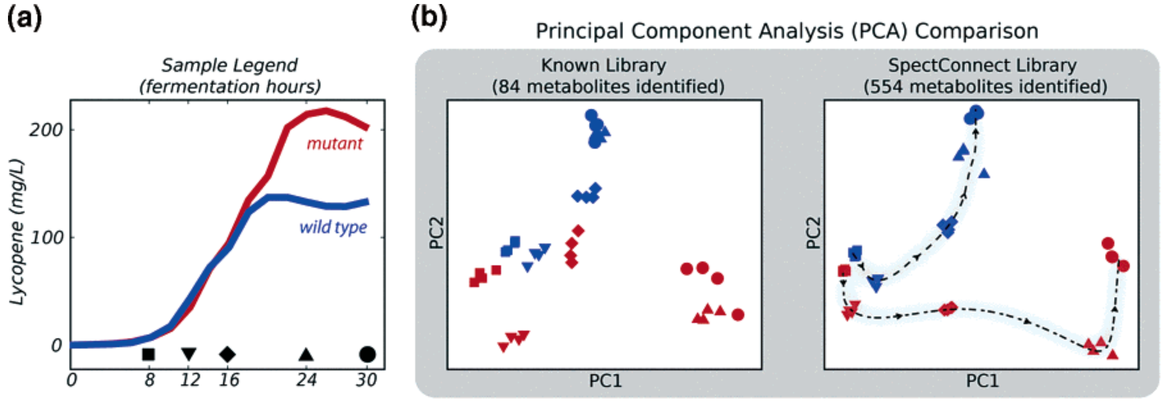

Figure 3. Including all conserved components in a PCA analysis better captures biological variation in metabolite profiles of cell physiological states. The first panel shows lycopene production as a function of time for the reference strain and the mutant strain (strain 1) in separate bioreactors. The 40 data points in the second and third panels represent identical experimental metabolic measurements from 40 samples. Four samples are taken from each strain at five different times, as denoted by the shapes on the *x*-axis in the first panel. (A) With the library of reference spectra, some separation is seen between strains, though some time points are difficult to resolve. We particularly note the lack of separation at time points 24 and 30 h. (B) After using SpectConnect to create a global library of conserved components, we obtain a better characterization of metabolic states and profiles. The SpectConnect library allows better resolution between the two strains and even suggests more distinct "trajectories" along sample times than is possible using the MCF reference library, including a divergence of trajectories before a difference in lycopene production is detectable.

performed in this work took reasonable computation times, ranging from minutes to a few hours.

Since our approach relies on adjustable thresholds at which two components are considered similar and thus conserved, some caveats related to our assumptions should be explicitly addressed. Using set thresholds for spectral similarity or similarity in retention time necessarily implies distinguishing sharply between similar cases on either side of a threshold. This distinction is obviously not ideal, and for these reasons an automated system can never fully replace an experienced and knowledgeable researcher. Overall, the ability of our algorithm to systematically track conserved components relies upon intelligently chosen assumptions, the choices of which inherently create finite resolution limitations to the exhaustiveness of the conserved component search. For instance, using the *E. coli* data set with 554 total metabolite peaks, changing the elution threshold will yield 549 to 569 metabolite peaks (at 2 and 0.5 min), changing the similarity threshold will yield 440 to 602 metabolite peaks (at 90% and 70% similarity), and changing the support threshold will yield 308 to 1659 metabolite peaks (at 100% and 50% required conservation within replicates). Overall, these variations seem reasonable or expected, especially considering the large magnitude of change in threshold parameters and the fact that some of the parameters (i.e., 50% required conservation within replicates) will intuitively yield a significant number of false positives. The more detailed parameter perturbation experiments in Supporting Information Table 2 support the fact that, at least in vitro, SpectConnect is robust with respect to identifying the known conserved and differential metabolites in a sample. In general, though, it is clear that the exact values chosen for thresholds have some effect on the method's results. Nonetheless, for the purposes of trying to track as many metabolite peaks as possible in as simple a fashion as possible, we believe that such a cost is marginal compared to the benefit of a broadened scope of analysis.

The selection of these thresholds is based upon experimental protocols and simple heuristics. The default similarity threshold of 80% is chosen to conform with the commonly implemented assumption that 80% similarity for comparison of a spectrum to a library represents a likely match. The default support threshold was chosen as 75% of samples, allowing for some experimental noise from the theoretical value of 100% but attempting to avoid inclusion of artifactual spectra that may occur at random. Since we used no internal retention index standards, we allowed for elution similarity thresholds of 1 min to account for column drift and other noise in retention-time data. The use of internal retention index standards would allow the reduction of elution time similarity thresholds, likely resulting in even finer resolution and less noise in SpectConnect's results. Finally, one may note that our preprocessing method for AMDIS data decreases the resolution of our approach. For instance, SpectConnect cannot distinguish between isomers with very similar elution times. Improved peak enumeration and deconvolution methods will help remove the need for this preprocessing and thus restore finer resolution to our method. As such, this shortcoming is representative of the current implementation and is not inherent to the general approach.

Despite these limitations, SpectConnect shows a great deal of promise for future applications. Removing the need for a library of reference spectra not only allows superior data analysis but also gives more flexibility in metabolomic sample preparations. Because SpectConnect works without adjustment on GC/MS data produced using any derivatization chemistry (including the most popular TMS-based techniques), the adoption of newer, potentially simpler derivatization methods[32] is made more practical, as the need for a comprehensive reference spectrum library is not as pressing. Just as chromatographic and mass spectrometry instrument parameters are explored and optimized,[11] variants of a derivatization chemistry may be tested to increase the range of measurable compounds.

When interesting metabolite peaks are identified using SpectConnect, established methods can be used to ascertain the molecular identities of components that cannot be matched to a library reference spectrum. Specifically, exact mass measurements from high-resolution mass spectrometers can be used to find molecular composition and structure; additional context clues like relative elution time and isotope ratios can further help pinpoint molecular identity. These methods have been adopted since the early stages of metabolomics research;[18] revision and refinement of these methods is still an active research area, as seen by recent efforts to enhance our ability to structurally identify high-molecular-weight metabolites.[33] Sharing these newly characterized spectra in recently created public mass spectrum databases[34] would enable even faster exploration of previously unidentified cellular metabolites. While we do not address such issues in this manuscript, pursuit of structural identification for unknown metabolites that we classified as highly discriminatory is certainly the next step in better characterizing the metabolic perturbations in our *E. coli* strains. Such identifications and characterizations may play a significant role in advancing the capabilities of the field of metabolomics.

SpectConnect also allows for the potential to thoroughly explore the differences between metabolomes of different species. By comparing many diverse species we may be able to determine the extent to which genetic similarity correlates with metabolomic similarity. That is, we can perform a more accurate assessment of the phylogenetic uniformity of the metabolome than is otherwise currently possible, as we will not be limited to those metabolites that are currently well-known and well-characterized.

In total, SpectConnect's direct contributions are significant: it helps broaden the scope of the systematic search for biomarkers, and it provides a unique, powerful, automated, and simple tool for interpreting complex, high-dimensional metabolomic data.

[32] Villas-Boas, S. G.; Moxley, J. F.; Akesson, M.; Stephanopoulos, G.; Nielsen, J. *Biochem. J.* **2005**, *388*, 669–677.

(33) Kind, T.; Fiehn, O. *BMC Bioinf.* **2006**, *7*, 234.
(34) Schauer, N.; Steinhauser, D.; Strelkov, S.; Schomburg, D.; Allison, G.; Moritz, T.; Lundgren, K.; Roessner-Tunali, U.; Forbes, M. G.; Willmitzer, L.; Fernie, A. R.; Kopka, J. *FEBS Lett.* **2005**, *579*, 1332–1337.
(35) Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol.* **2000**, *18*.
(36) Roessner, U.; Willmitzer, L.; Fernie, A. R. *Plant Physiol.* **2001**, *127*.
(37) Taylor, J.; King, R. D.; Altmann, T.; Fiehn, O. *Bioinformatics* **2002**, *18* (Suppl. 2).
(38) Roessner-Tunali, U.; Hegemann, B.; Lytovchenko, A.; Carrari, F.; Bruedigam, C.; Granot, D.; Fernie, A. R. *Plant Physiol.* **2003**, *133*.
(39) Verdonk, J. C.; Ric de Vos, C. H.; Verhoeven, H. A.; Haring, M. A.; van Tunen, A. J.; Schuurink, R. C. *Phytochemistry* **2003**, *62*.
(40) Prithiviraj, B.; Prithiviraj, A. V.; Kushalappa, A. C.; Yaylayan, V. *Eur. J. Plant Pathol.* **2004**, *110*, 371–377.
(41) Duran, A. L.; Yang, J.; Wang, L.; Sumner, L. W. *Bioinformatics* **2003**, *19*.
(42) Morris, C. R.; Scott, J. T.; Chang, H.-M.; Sederoff, R. R.; O'Malley, D.; Kadla, J. F. *J. Agric. Food Chem.* **2004**, *52*.
(43) Barsch, A.; Patschkowski, T.; Niehaus, K. *Funct. Integr. Genomics* **2004**, *4*.
(44) Strelkov, S.; von Elstermann, M.; Schomburg, D. *Biol. Chem.* **2004**, *385*.
(45) Vikram, A.; Prithiviraj, B.; Kushalappa, A. C. *J. Plant Pathol.* **2004**, *86*, 215–225.
(46) Bino, R. J.; Ric de Vos, C. H.; Lieberman, M.; Hall, R. D.; Bovy, A.; Jonker, H. H.; Tikunov, Y.; Lommen, A.; Moco, S.; Levin, I. *New Phytol.* **2005**, *166*.
(47) Devantier, R.; Scheithauer, B.; Villas-Boas, S. G.; Pedersen, S.; Olsson, L. *Biotechnol. Bioeng.* **2005**, *90*.
(48) Kaplan, F.; Kopka, J.; Haskell, D. W.; Zhao, W.; Schiller, K. C.; Gatzke, N.; Sung, D. Y.; Guy, C. L. *Plant Physiol.* **2004**, *136*.
(49) Desbrosses, G. G.; Kopka, J.; Udvardi, M. K. *Plant Physiol.* **2005**, *137*.
(50) Tarpley, L.; Duran, A. L.; Kebrom, T. H.; Sumner, L. W. *BMC Plant Biol.* **2005**, *5*.
(51) Diran Herebian, B. H.; Marner, F.-J. *Metabolomics* **2005**, 1–8.
(52) Tikunov, Y.; Lommen, A.; Ric de Vos, C. H.; Verhoeven, H. A.; Bino, R. J.; Hall, R. D.; Bovy, A. G. *Plant Physiol.* **2005**, *139*.
(53) Villas-Boas, S. G.; Moxley, J. F.; Akesson, M.; Stephanopoulos, G.; Nielsen, J. *Biochem. J.* **2005**, *388*.
(54) Brosche, M.; Vinocur, B.; Alatalo, E. R.; Lamminmaki, A.; Teichmann, T.; Ottow, E. A.; Djilianov, D.; Afif, D.; Bogeat-Triboulot, M.-B.; Altman, A.; Polle, A.; Dreyer, E.; Rudd, S.; Paulin, L.; Auvinen, P.; Kangasjarvi, J. *Genome Biol.* **2005**, *6*.

ACKNOWLEDGMENT

We acknowledge Jesper Højer-Pedersen for helpful discussions and effort in framing the problem; Hal Alper, Joanne Kelleher, Curt Fischer, and Jose Aleman for thorough reviews of the manuscript; Silas Villas-Bôas for the original MCF library and critical review of the manuscript; and the NSF Graduate Research Fellows Program (MPS and LVT) and NIH Metabolomics Roadmap Initiative, DK070291, for financial support. The first two authors contributed equally to this work.

SUPPORTING INFORMATION AVAILABLE

Experimental methods, statistical methods, supplementary tables. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review August 9, 2006. Accepted October 25, 2006.

AC0614846