XCMS[2]: Processing Tandem Mass Spectrometry Data for Metabolite Identification and Structural Characterization

H. P. Benton, **D. M. Wong**, **S. A. Trauger**, and **G. Siuzdak**[*]
Department of Molecular Biology and The Center for Mass Spectrometry, The Scripps Research Institute 10550 North Torrey Pines Road, La Jolla, California 92037

Abstract

Mass spectrometry based metabolomics represents a new area for bioinformatics technology development. While the computational tools currently available such as XCMS statistically assess and rank LC–MS features, they do not provide information about their structural identity. XCMS[2] is an open source software package which has been developed to automatically search tandem mass spectrometry (MS/MS) data against high quality experimental MS/MS data from known metabolites contained in a reference library (METLIN). Scoring of hits is based on a "shared peak count" method that identifies masses of fragment ions shared between the analytical and reference MS/MS spectra. Another functional component of XCMS[2] is the capability of providing structural information for unknown metabolites, which are not in the METLIN database. This "similarity search" algorithm has been developed to detect possible structural motifs in the unknown metabolite which may produce characteristic fragment ions and neutral losses to related reference compounds contained in METLIN, even if the precursor masses are not the same.

Liquid chromatography mass spectrometry (LC–MS) combined with comprehensive quantitative data analysis has allowed LC–MS to become the primary platform for metabolic profiling experiments.[1,2] However, even as "shotgun" proteomic data analysis has evolved over the past decade,[3] metabolomics is still in the beginning stages of its' informatic development. Metabolomics experiments are traditionally targeted, where specific metabolites are quantified, and untargeted metabolomics experiments include relative quantification of all observed metabolites. We have termed the combination of these strategies "sniper" metabolomics, where initial LC–MS metabolite profiles are followed by LC–MS/MS experiments on selected, highly differentiated targets. The resulting output from a typical "sniper" metabolomics experiment is the statistical evaluation of the most significantly changing metabolites followed by their structural identification.

A common goal of most bioinformatic platforms in metabolomics is to allow users to find and statistically assess features that show significant change between sample groups. The most significant features are then selected for targeted tandem MS (MS/MS) analysis, thereby guaranteeing high quality MS/MS data which provides high confidence to data interpretation. Recently, three different open source MS based metabolite profiling software packages were released including XCMS,[1] MathDAMP,[4] and Met-IDEA.[5] Met-IDEA and MathDAMP both have differential analysis capabilities and present useful platforms for GC and LC–MS analysis. MathDAMP takes advantage of the language Mathematica to present data in a variety of graphical representations. Met-IDEA is written in Microsoft's ".NET" language and features a simple, user-friendly interface that accepts capillary electrophoresis mass spectrometry data. Instrument manufacturers have also produced software packages for metabolomics including

* To whom correspondence should be addressed. E-mail: siuzdak@scripps.edu. Fax: (858) 784-9496.

Water's MarkerLynx, AB Sciex's MarkerView and Agilent's Genespring-MS. MarkerLynx and MarkerView utilize principal component analysis (PCA) for differential comparison and visualization of data sets and include a *t* test for evaluating significant differences. MarkerLynx also takes advantage of Waters' ultra performance LC system for spectral alignment. Genespring-MS, which has incorporated aspects of XCMS, offers an array of statistical tools for data postprocessing. XCMS can be distinguished from these programs as it allows for nonlinear spectral alignment and differential analysis as well as being able to run on all UNIX, Apple's OS X, and Microsoft Windows systems. Distributed through Bioconductor,[6] under the General Public License or GNU open source, XCMS is written in "R"[7] allowing it to be highly accessible and modular.

The success of these bioinformatic platforms is based on their ability to identify, quantify, and statistically assess features of interest. The next step in data analysis is the structural characterization of metabolites representing these features. This is commonly the most time-consuming aspect of metabolomics.[2,8] To accomplish this, MS/MS experiments are used to ascertain structural information about the molecule. Experimentally obtained fragmentation data can be compared with a reference MS/MS data set to identify the structure. Among the MS/MS databases available for metabolite identification are NIST MS/MS,[9] MassBank,[10] LipidMAPS,[11] and METLIN. NIST MS/MS and LipidMAPS use ion trap technology for their MS/MS data library. MassBank uses data generated by a combination of triple-quad, Q-TOF, and ion trap mass spectrometers. The METLIN[12] database has been recently expanded to include mass spectra from over 14 000 endogenous, exogenous, and theoretic di/tripeptide metabolites with the addition of their publically available q-TOF MS/MS data. To date, one commercial package, HighChem,[13] automates the analysis of MS/MS libraries and fragmentation mechanisms for the analysis of metabolite structure. However, its algorithm is not yet open to inspection or modification. Alternatively, manual comparative MS/MS analysis is available from databases such as METLIN, NIST, LipidMaps, and Mass-Bank. As more metabolite databases are starting to include MS/MS data, adding the capability to perform searches against this data is a natural progression of XCMS development.

Here we report XCMS[2], an open source next generation metabolomics platform. XCMS[2] is an extension of XCMS, as it features the same reliable peak picking, alignment, statistical analysis of features but with the added capability of automatic searching of MS/MS spectra against the METLIN database. Included in this platform is a "similarity search" functionality that analyzes MS/MS patterns for structural classification of unknown molecules.

Experimental Section

Materials and Methods

HPLC grade acetonitrile, water, and methanol and were obtained from Fisher Scientific. Formic acid, clarithromycin, oleamide, aminoantipyrine, omeprazole, benzyl-amine, reserpine, salbutomal, glycerophosphocholine, sulfanilamide, and 1-palmitoyl-*sn*-glycero-3-phosphocholine, and male serum were obtained from Sigma Chemical Co. (St. Louis, MO) in high purity. An Agilent 6510 ESI Q-TOF was used to generate the LC–MS and LC–MS/MS data for XCMS and XCMS[2] analysis.

Metabolite Identification Using Chemical Standards

To test the efficiency of the algorithm, a 1 μM stock solution with 10 standards were analyzed in positive (+) ion mode via LC-MS/MS analysis. All 10 compounds were diluted with 50:50 (v/v) methanol/water and separated with a Zorbax SB-C18 column (5 μm particle size, 150 mm $\times$ 0.5 mm, Agilent) with an injection volume of 8 μL. A flow rate of 12 μL/min with elution buffers of A, water with 0.1% formic acid, and B, acetonitrile with 0.1% formic acid. The total

LC–MS analysis time was 37 min. The initial gradient was 5% B for the first 2 min and continued with percentage and time range as follows: 95% B at 32 min, 5% B at 37 min. MS/MS mass spectral data were stored in profile data mode. A m/z scan range of 50–900 was used for the analysis. The precursor ions of all 10 compounds were subjected to data-dependent MS/MS at a "preferred ion setting" of 200 ppm and a "medium (~4 m/z)" isolation window. Two reference ions with m/z of 121.0509 and 922.0098 were used as internal mass calibrants. The Agilent 6510 ESI Q-TOF MS data file (.d) was converted with centroiding and deisotoping to a file format suitable for XCMS[2] (mzData). This data was processed by XCMS[2] software which compared the experimental fragmentation data to the MS/MS data in METLIN (metlin.scripps.edu).

Plasma Extraction and Preparation

The starting volume of serum was 50 μL. Serum was extracted with cold methanol for protein removal. The sample was briefly vortexed for 10 s and incubated at −20 °C for 20 min. After incubation, the extracted serum was centrifuged at 14 000 g for 10 min at 4 °C. The supernatant was dried, resuspended in 95:5 (v/v) acetonitrile/water, and separated into two aliquots (~25 μL). In the first aliquot, 5 μL of 1-palmitoyl-sn-glycero-3-phosphocholine was spiked into the extracted serum sample. In the second aliquot, 5 μL of methanol was added to keep the concentration consistent. Both aliquots were vortexed for 1 min and transferred to HPLC vials for LC-MS and LC-MS/MS analysis.

Spiked and Unspiked Plasma Analysis

Both serum samples (spiked and unspiked) were separated using a reverse-phase HPLC Zorbax SB-C18 column (5 μm particle size, 150 mm × 0.5 mm, Agilent) and analyzed by Agilent 1100 LC/MSD system for a total analysis time of 60 min. The column flow rate was set at 12 μL/min with an injection volume of 8 μL. The elution buffers were A, water with 0.1% formic acid, and B, acetonitrile with 0.1% formic acid. The gradient started at 5% B for 5 min and continued with a percentage of B and time as follows: 95% B at 55 min, 5% B at 60 min. The data was collected in positive ion mode with a m/z range of 50–1500. Spiked and unspiked samples were analyzed in triplicate. Two methanol washes were performed in between sample analysis to prevent sample carryover. The raw files (.wiff) were converted to mzData format and analyzed by XCMS (version 1.10.9). The parameters were optimized for the samples using fwhm = 38, profmethod = "binlinbase", step = 0.05, steps = 3, and span = 0.9. All other XCMS parameters remained at default settings. After XCMS analysis, the difference report was filtered by the p-value, intensity, and fold difference. The features with the lowest p-value, intensities above 2000, and highest fold difference were analyzed by targeted LC–MS/MS analysis.

For the targeted LC–MS/MS analysis, the HPLC column, reference mass corrections, elution solvent, and elution gradient were identical to the LC–MS analysis mentioned above. The only difference with this LC–MS/MS analysis is the data storage and mass ranges. The MS and MS/MS product ion spectra were both collected at a range of 100–1500 m/z in positive ion mode. Targeted MS/MS analysis was performed using a narrow isolation window of 1.3 m/z, a collision energy of 20 V, and a retention time window of 1 min. Data was stored in centroid mode, and all precursor ions were targeted with a collision energy of 20 V. This data was converted to the "mzData" file format and was analyzed using both sets of computational methods against the METLIN database.

MS/MS Database in METLIN

All the MS/MS data on standards generated for METLIN were acquired on an Agilent 6510 ESI Q-TOF by direct infusion and varying collision energies (0, 10, 20, and 40 V).

Theory, Results, and Discussion

MS/MS data is extensively used with Mascot,[14] X-Tandem,[15] and Sequest[16] proteomic software packages where the experimentally obtained MS/MS data are compared against theoretical fragment ions. These search engines are able to generate a list of theoretical fragments from peptides based on the rules governing peptide fragmentation via collision induced dissociation (CID). However, predicting the fragmentation of endogenous metabolites is computationally challenging due to the complexity of their structures and difficulty of calculating energetically favorable losses for larger molecules.

With the "sniper" metabolomics approach, metabolite features are statistically tested for their quantitative significance, and once a statistically interesting feature is identified, structural identification is the next stage in the discovery process. XCMS[2] has been designed to allow users the full alignment and statistical capabilities of XCMS, in addition to the newly added capability of searching MS/MS data for the purposes of metabolite identification.

Preprocessing

The preprocessing step is designed to increase the accuracy of the MS/MS spectra by processing it through a signal-averaging algorithm. The spectra are sorted in the chromatographic domain with a user defined signal-to-noise ratio cutoff and a retention time window into which the spectra are sorted. Equation 1 calculates a noise level and uses only mass spectral data that is greater than calculated noise. Where, "x", a user defined signal-to-noise ratio, $\bar{I}$ is defined as the mean intensity of the current spectra, gap is the distance between each adjacent peak with isotopic peaks excluded, MZ_r is the mass range of the spectra (max−min), and I_{min} is the minimum intensity of the spectra. After this calculation, the spectra are collected and the data is indexed by precursor mass for quick referencing.

$$\text{signal} > x \left(\frac{I\left(1 - \frac{\text{gap}}{MZ_r}\right) + I_{min}\left(\frac{\text{gap}}{MZ_r}\right)}{1 - \frac{\text{gap}}{MZ_r} + \frac{\text{gap}}{MZ_r}} \right)$$

(1)

Matching/Spectral Alignment

There is still discussion in the MS community about the optimum means of assigning a confidence value to the similarity between two spectra. For metabolomics, the current approaches include the "shared peak count"[17] and the "spectral convolution"[18] methods, although there are variations on both of these methods. The "shared peak count" is the most common matching system and is calculated based on the number of masses in common between reference and experimental spectra. "Spectral convolution" processes the convolution of the reference spectra against the experimental spectra, and the results are demonstrated through a difference matrix. The difference matrix is made up of the differences between each successive peak. A difference of zero equals a match. A known difference, such as a mass defect corresponding to a phosphate group, is also considered a match. This method allows for quick identification of altered metabolites and more robust matching; however, it is more computationally time-consuming. Accurate mass measurements improve the performance of these algorithms, as the search space is typically reduced to only a few metabolites. For example a search in the METLIN database for a mass of 258 ± 1 Da generates 24 molecules. If the accuracy of the measurement is increased to ± 10 ppm (ppm) accuracy, 5 molecules are retrieved. However, if we have accurate mass measurements at the 5 ppm level, a search for 258.1101 ± 5 ppm returns only one molecule. The utility of accurate mass measurements in the identification of small molecules have been shown throughout the literature,[19] although

MS/MS data along with retention time comparison with chemical standards is typically required for structural identification.

The XCMS[2] algorithm employs a modification of the "shared peak count" method, which takes the precursor mass from the MS scan and uses an indexed METLIN file for removal of any nonmatching precursor mass. After narrowing the search space by precursor mass, XCMS[2] matches the collision energy of the MS/MS spectra. If there is no current matching collision energy in the experimental spectra, the algorithm will attempt to match with a higher collision energy spectrum. Use of higher collision energy spectra gives the best performance, which will be discussed in more detail in the "similarity search" section. The software then performs comparative analysis between fragment data and a reference spectrum. The match is made using a window of user specified error in parts per million (ppm) for each fragment mass. If the masses are outside the error window, then the algorithm moves onto the next fragment mass. This operation is performed using two different matrices, a similarity and distance matrix. Each cell in the matrix is calculated using eq 2 and the distance matrix using eq 3. The similarity matrix "M" calculates a score "$S_{i,j}$" which defines the similarity between two MS/MS spectra.

$$S_{ij} = \max(M_{i-1j}, M_{ij-1}, M_{i-1j-1}) - C$$

(2)

In the matrix "M", "i" and "j" are positions, where "i" is the horizontal coordinate and "j" is the vertical coordinate. For each cell in the matrix, a score is calculated ($S_{i,j}$). If the two compared masses are a match, then "C" the cost is equal to 0. However, if the two masses are outside the designated ppm error window, then the cost is equal to 1. The score that is placed into the similarity matrix at $S_{i,j}$ is a maximum of the surrounding matrix cells, the cell above ($M_{i,j-1}$), the cell to the left ($M_{i-1,j}$), and the cell directly diagonally above ($M_{i-1,j-1}$) minus the cost value, which is either 0 or 1. In the similarity matrix, the starting score is equal to the maximum number of fragments of the two spectra or what would be the maximum similarity. As the algorithm progresses, the matrix is filled up by the score values. The final score is always the bottom right cell in the matrix, which denotes the amount of similarity. The distance matrix also calculates a score defining the likeness between the two MS/MS spectra. The initial score $D_{i,j}$, (when $i = 1$ and $j = 1$) starts as "0" in the difference matrix and is processed in accordance to eq 3.

$$D_{ij} = C + \min(M_{i-1j}, M_{ij-1}, M_{i-1j-1})$$

(3)

These two scoring systems for calculating D and S are a good measure of the similarity between both spectra. However, D or S by themselves do not represent the whole picture. In many cases, the experimental and reference spectra may not have the same number of peaks. A second equation can be used to incorporate both scores into one system. Although many different scoring functions have been suggested for spectral matching,[17,20–22] here a single scoring system is used to generate a score so that all the results can be measured equally. By taking both the similarity and the distance scores, a measure of how good the match is can be formulated. This score is useful to have as a percentage match.

$$H = \min_{\text{Length}}(\text{Exp}, \text{Ref}) - \text{diff}_{\text{Length}}(\text{Exp} - \text{Ref})$$

(4)

$$L = - \max_{\text{Length}} (\text{Exp}, \text{Ref}) \tag{5}$$

$$S_\% = 100 \times \frac{(S - D) + L}{L - H} \tag{6}$$

Equation 4 represents what the score would be if the two spectra were identical (*H*). The "*H*" term is calculated by $S - D$ when the spectra are 100% identical. The *S* term is the minimum number of comparable objects (number of fragments or neutral losses) or $\min_{\text{length}}(\text{Exp}, \text{Ref})$ from either the experimental spectra Exp or the reference spectra Ref. This is then subtracted against distance matrix score (*D*) when the spectra are identical which can be found by the difference between the number of compa-rables of the two spectra $\text{diff}_{\text{length}}(\text{Exp}, \text{Ref})$. Equation 5 is the score when the two spectra are completely different (*L*). In this case the similarity score would be equal to zero by definition. If the two spectra are totally different then there is no similarity. Finally the *D* term is represented by $\max_{\text{length}}(\text{Exp}, \text{Ref})$ which is the difference between the two spectra in terms of the number of fragments they have.

The similarity score *S* is subtracted from the difference score *D*. This value is then added to *L* to change the score to the scaled score. This value is then divided by *L* minus *H* and made into a percentage. The result is the percentage match or $S_\%$. This value has a normalization function, allowing cross compound comparison. However, if a match is not found, then the molecule is either not in the database or the ppm error window needs to be increased.

The above method is illustrated in Figure 1 showing the operation of design. It shows the flow from a global analysis using XCMS to a targeted analysis using XCMS[2].

XCMS[2] Positive Identification

To test the search algorithm, a simple mixture of 10 compounds from the METLIN database were analyzed via LC–MS/MS. Of the 10 compounds, 9 are ionized in the positive mode, with the exception of sulfanilamide. Table 1 shows the molecules which were correctly identified by the algorithm with a high percentage score. Omeprazole was outside of the ppm error range set for precursor masses and therefore it did not appear to match but was identified by performing a "similarity search" discussed later. Albuterol's score is low due to the low quality of its MS/MS spectrum. The data-dependent behavior of the mass spectrometer picks ions that are of the greatest intensity in a particular scan. In this case, even with a preference list, other molecules were chosen for MS/MS while albuterol was eluting off the column and therefore albuterol did not have sufficient abundance for MS/MS analysis. This score would improve if MS/MS was done on the compound while it was at its peak eluting time. Additionally, the score is affected by the ppm error window.

Like XCMS, the XCMS[2] algorithm provides a visualization tool for the side-by-side comparison of the reference and experimental spectra. A comparison plot of the two MS/MS spectra facilitates visual confirmation. As can be seen in Figures 2 and 3, the relative intensities can vary due to different acquisition parameters such as collision energies and times, and although not shown, absolute intensity can also vary with analyte concentration. Originally, METLIN's MS/MS data was an average of spectra collected at four different collision energies: 0, 10, 20, and 40 V. However, testing the algorithm on the averaged data set gave lower score values than expected. This was due to the inclusion of smaller and larger fragment ions from higher and lower collision energies, respectively. Averaging the spectra gave a fictitious spectrum which consequently did not match any single collision energy spectrum. In a typical

experiment, a single collision energy value is chosen for an ion. If the experimental spectrum is not at one of the METLIN defined collision energies, then the software will choose the closest corresponding energy. It is worth noting that the $XCMS^2$ algorithm currently ignores intensity data for matching. However, it does require an intensity threshold.

Our experience has shown that with this method, the MS/MS pattern of a single collision energy spectrum against a reference spectrum is specific enough to correctly identify the compound, even without the use of accurate mass on the precursor ion. The use of accurate mass enhances the speed of the algorithm and gives another level of conformation. Figure 4 shows how structural information provided by the MS/MS data removes ambiguity when two molecules have the same accurate mass.

$XCMS^2$ Similarity Search

When the complete reference database with an unrestricted precursor mass is searched against, it was observed that along with an accurate hit on the correct compound many other compounds had a reasonable score, as demonstrated in Figure 5. Upon further investigation, these compounds were found to be structurally related leading to the idea of a "similarity search". With further analysis of both the experimental and reference data, common structural elements can be identified, providing useful information for identification.

Using the previously described scoring schema for compound comparison, the spectra are analyzed for similarity to the reference MS/MS spectrum, either by neutral losses or fragment ions. Both the neutral loss search and fragment ion search work on the same basis as the previously described scoring schema except the neutral loss search computes the difference in mass between each adjacent peak. If there is a high degree of similarity, these compounds are reported in a table of results. While the amount or percentage of similarity is defined by the user, we recommend using a similarity value of greater than 70%. The results table, seen in Table 2, contains important information on the compounds, including neutral losses and a fragment score, as well as the common structural motif identified. Table 2 is a condensed version of the two reported tables from the "similarity search" where the most significant hits are reported and figures have been added for clarity. The first reported tables hold information on common structural motifs and collision energies that were used for the match. It was found that for a MS/MS search, using a collision energy that is greater or equal to the experimental value yields the strongest match. This phenomenon can be seen in Figure 3 and is believed to occur due to systematic variations in the concentrations of metabolites and intensities. The second table provides a quick overview of the mass and gives a common collected substructure. This is obtained by a frequency counter which looks at all the common neutral losses and fragment ions separately, from the first table. The second table can be very useful for a comparison of the experimentally obtained spectrum to reported molecules.

The algorithm was tested using the same mixture, with the result correctly identifying the compounds. The results in Table 2 show how each molecule has a high fragment score, however the neutral loss score can be very low. This is due to the change in small fragmentation peaks that shift the neutral losses. Any ions coisolated and cofragmented with the precursor will dramatically decrease the accuracy of the neutral loss search. The neutral loss score can also be low because not every possible neutral loss is calculated; therefore, if a molecule has a fragment ion in the neutral loss region, it will not be seen.

$XCMS^2$ Analysis of LC–MS/MS Plasma Data

With the use of data from the spiked plasma extract experiment, both quantitative analysis reports and compound identification was performed using $XCMS^2$, as is demonstrated in Figure 1. XCMS results show elevated levels of the spiked metabolite.

To confirm the structure of the spiked metabolite, a separate data dependent MS/MS run with a preference list was performed on the sample. From the XCMS report, the top 200 results were taken and a filter applied so that intensities above 5000 counts were selected. XCMS[2] correctly identified the spiked metabolite as 1-palmitoyllysophosphatidylcholine. Other selected endogenous metabolites which were also identified with a high confidence including urocanic acid and methysergide.

Table 3 presents the similarity search results where these data were used with accurate mass searching to identify more compounds. Searching accurate mass alone of the unidentified molecules resulted in multiple identifications. With the similarity search, these were narrowed down to just a few molecules. The compound at m/z 510 and 516 Da are believed to be glycerophospholipid according to the similarity search from XCMS[2]. With the use of fragmentation and neutral loss scores, the list of accurate mass matches can be narrowed down to three molecules, since the similarity search suggests that the compound is a glycerophospholipid class. These three molecules can then be easily tested in a side-by-side analysis to determine the definite structures. While we recommend the use of a similarity value of greater than 70% for confident assignment of structural elements for an unknown metabolite similarity, using lower values can often provide useful clues about the actual structure.

Conclusion

XCMS's output report allows straightforward method creation for either targeted or preferred ion MS/MS lists, which are ideal for the "sniper" metabolomics approach. From this method, an XCMS[2] search can be used as a data processing tool for identifying metabolites of interest and as a data comparison tool with the "similarity search". Different databases can also be used with XCMS[2] by converting the database to the Metlin XML format and changing the address link to the new database. Future developments of XCMS[2] will include the use of basic structure–specific fragmentation rules and algorithms like the "similarity search" to facilitate structural elucidation of novel and yet unknown metabolite structures.

Acknowledgments

References

1. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. Anal Chem 2006;78:779–787. [PubMed: 16448051]
2. Want EJ, Nordstrom A, Morita H, Siuzdak G. J Proteome Res 2007;6:459–468. [PubMed: 17269703]
3. Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I, Mohammed S, Deery MJ, Howard JA, Dunkley T, Aebersold R, Kell DB, Lilley KS, Roepstorff P, Yates JR 3rd, Brass A, Brown AJ, Cash P, Gaskell SJ, Hubbard SJ, Oliver SG. Nat Biotechnol 2003;21:247–254. [PubMed: 12610571]
4. Baran R, Kochi H, Saito N, Suematsu M, Soga T, Nishioka T, Robert M, Tomita M. BMC Bioinf 2006;7:530.
5. Broeckling CD, Reddy IR, Duran AL, Zhao X, Sumner LW. Anal Chem 2006;78:4334–4341. [PubMed: 16808440]
6. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J. GenomeBiology 2004;5:R80. [PubMed: 15461798]

7. Team, RDC. The R Foundation for Statistical Computing; Vienna, Austria. 2007;

8. Dettmer K, Aronov PA, Hammock BD. Mass Spectrom Rev 2007;26:51–78. [PubMed: 16921475]

9. NIST, 2005.

10. Horai, H.; Suwa, K.; Arita, M.; Nihei, Y.; Nishioka, T. 55th ASMS Conference on Mass Spectrometry and Allied Topics; June 3–7, 2007; Indianapolis, IN.

11. Zemski Berry KA, Murphy RC. Anal Biochem 2006;349:118–128. [PubMed: 16384548]

12. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G. Ther Drug Monit 2005;27:747–751. [PubMed: 16404815]

13. Mistrik R. ThermoFinnigan. 2002

14. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Electrophoresis 1999;20:3551–3567. [PubMed: 10612281]

15. Craig R, Beavis RC. Bioinformatics 2004;20:1466–1467. [PubMed: 14976030]

16. Eng JK, McCormack AL, Yates JR III. J Am Soc Mass Spectrom 1994;5:976–989.

17. Pevzner PA, Mulyukov Z, Dancik V, Tang CL. Genome Res 2001;11:290–299. [PubMed: 11157792]

18. Pevzner PA, Dancik V, Tang CL. J Comput Biol 2000;7:777–787. [PubMed: 11382361]

19. Clauser KR, Baker P, Burlingame AL. Anal Chem 1999;71:2871–2882. [PubMed: 10424174]

20. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. J Proteome Res 2004;3:958–964. [PubMed: 15473683]

21. Eriksson J, Chait BT, Fenyo D. Anal Chem 2000;72:999–1005. [PubMed: 10739204]

22. Havilio M, Haddad Y, Smilansky Z. Anal Chem 2003;75:435–444. [PubMed: 12585468]
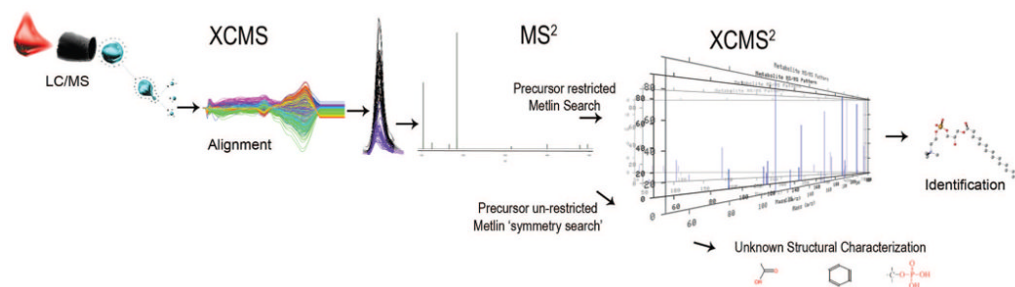
Figure 1.
The general workflow of XCMS and XCMS^2 employing the "sniper" approach: a single feature is found with statistical confidence and selected for MS/MS. The data from the MS/MS is then put through XCMS^2 and the feature is structurally identified using the METLIN database.
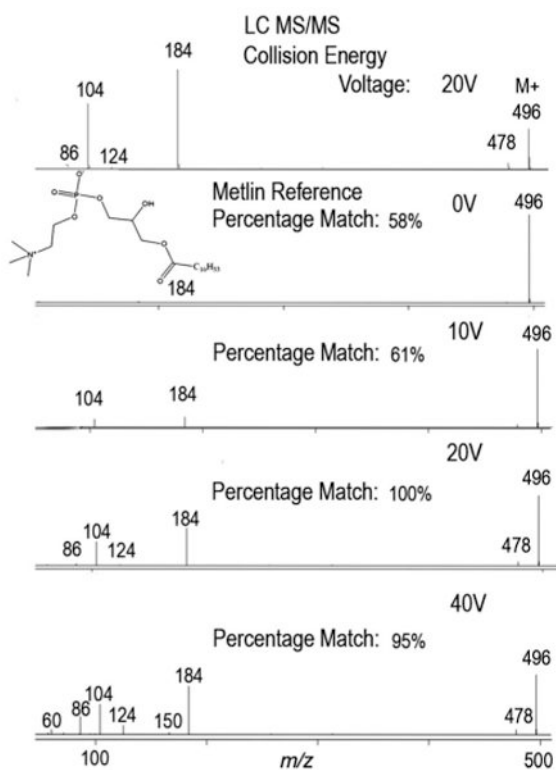
Figure 2.
The comparison of the different matches at the reference collision energy. As the voltage increases, more low mass fragment ions are generated. Using a collision energy that is close to the experimentally used value gives a high matching score.
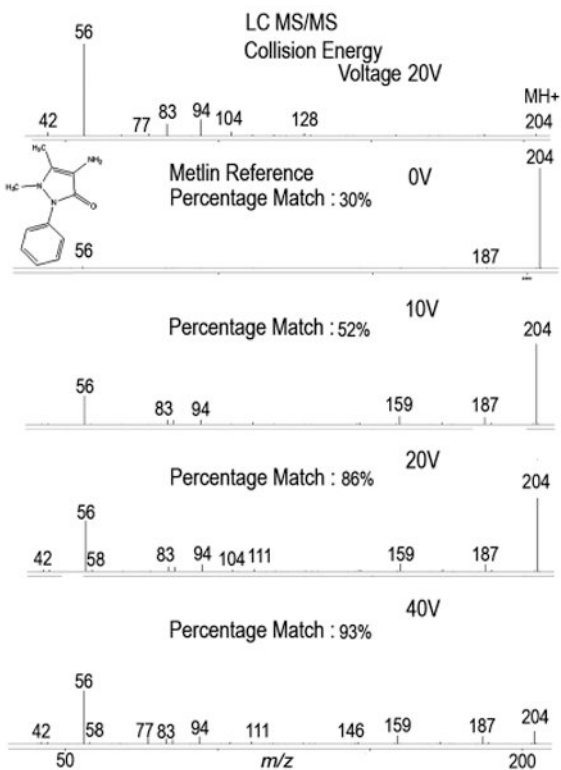
Figure 3.
The difference between the collision energies is seen again. However, the figure also shows how high collision energy spectra can match lower collision energy spectra due to signal-to-noise levels of low abundance peaks.
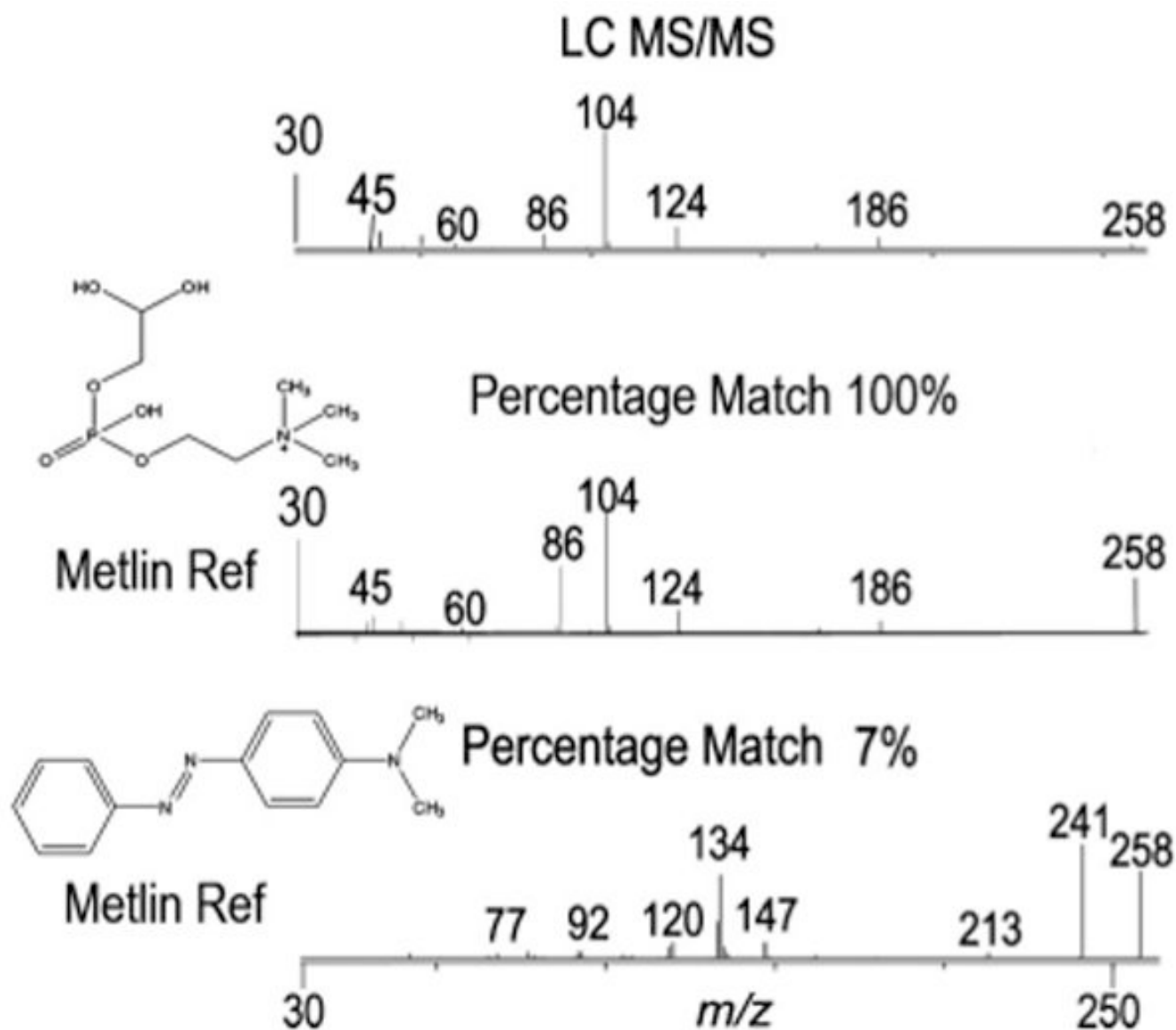
Figure 4.
A METLIN search using accurate mass only can yield multiple hits. However, searching with MS/MS data, glycerophosphocholine can be unambiguously identified.
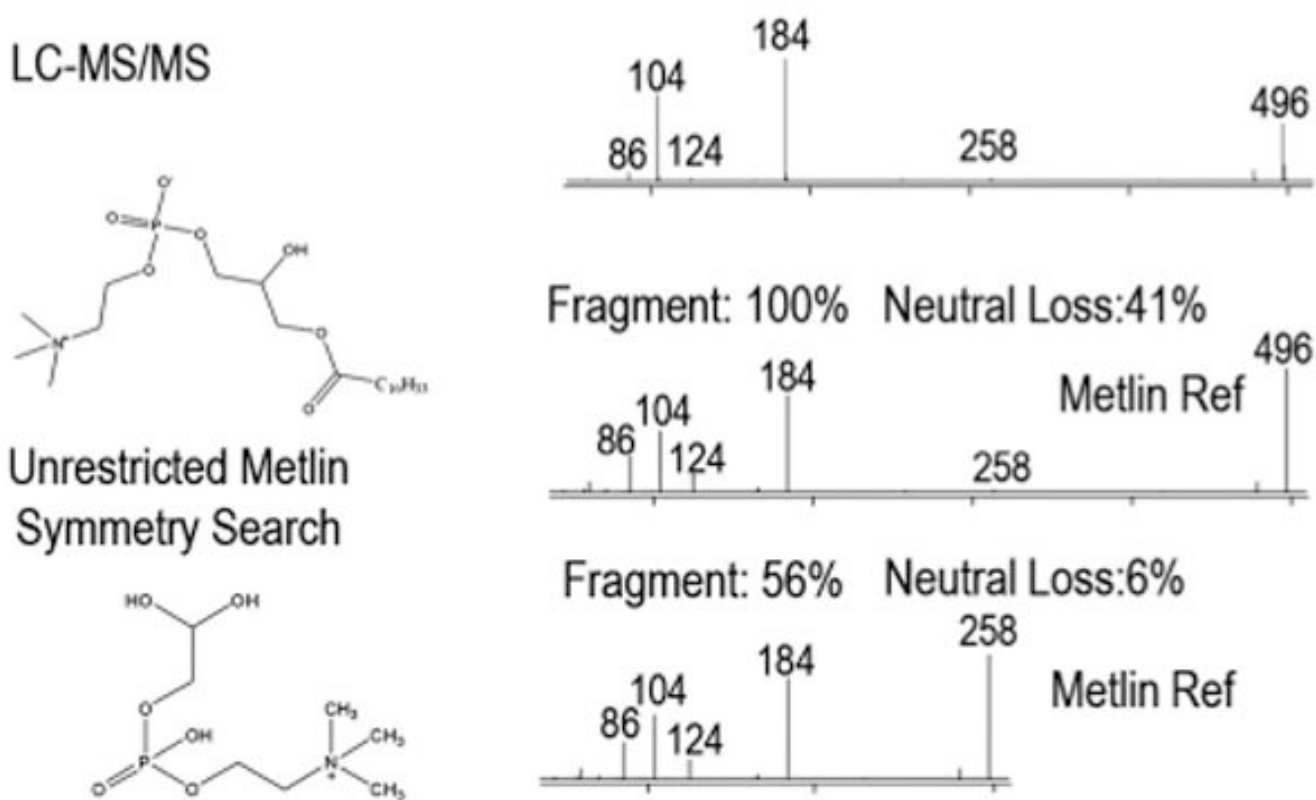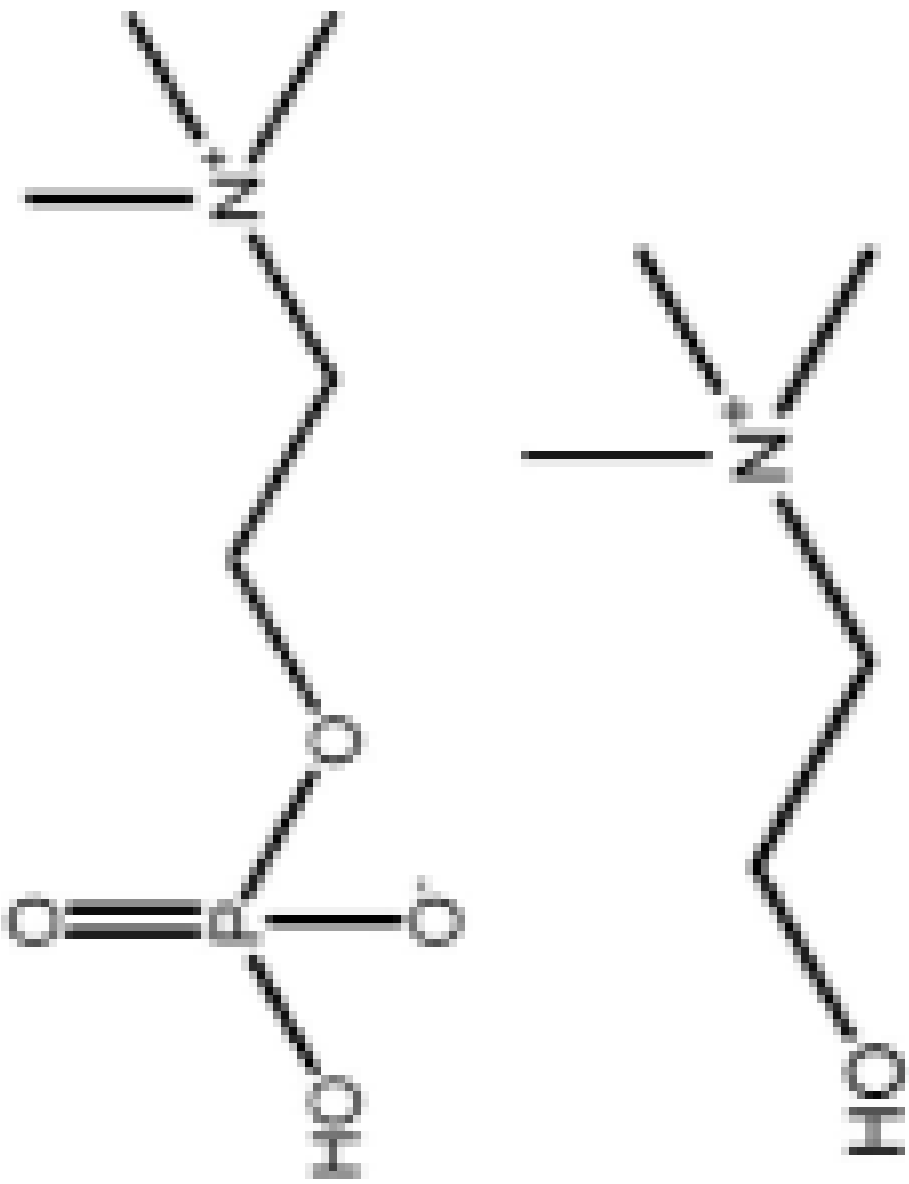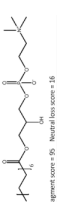
Figure 5.
Similarity search. The first spectrum is the MS/MS data collected from an experiment of 1-palitylophosphocholine. The spectra below show similarity to the unknown.

Table 1

Results for the Standard Molecule Identification[a]

		fragment ions			
identified molecule	precursor *m/z*	METLIN	matching	nonmatching	score
benzylazanium	108.082	11	8	10	75
glycerophosphocholine	258.113	21	14	7	95
albuterol	240.149	10	5	9	50
aminoantipyrine	204.155	44	19	25	79
omeprazole	**346.165**	**26**	**18**	**8**	**92**
clarithromycin	748.492	48	16	32	76
reserpine	609.300	139	41	98	69
1-palmitoyllysophosphatidylcholine	496.354	15	12	4	96
oleamide	282.286	67	29	38	78

[a]Identified molecules in bold are the METLIN search results that were inside the searched ppm error window, the other identified molecules are from the similarity search. Albuterol has a low score due to unoptimized LC–MS/MS parameters.
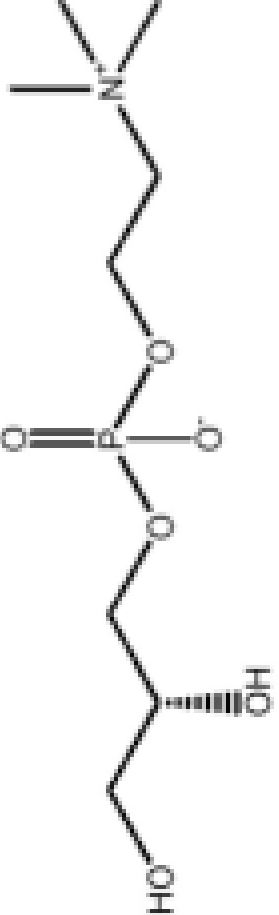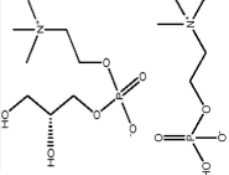
Identified common structure

Related molecule(mass)

Identified common structure

Related molecule(mass)

erythromycin A (733 Da)
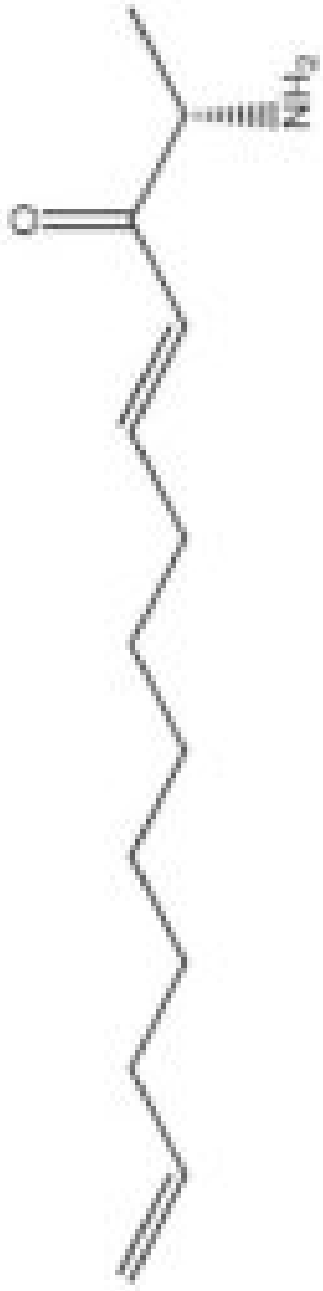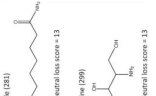
Fragment score = 75   Neutral loss score = 18
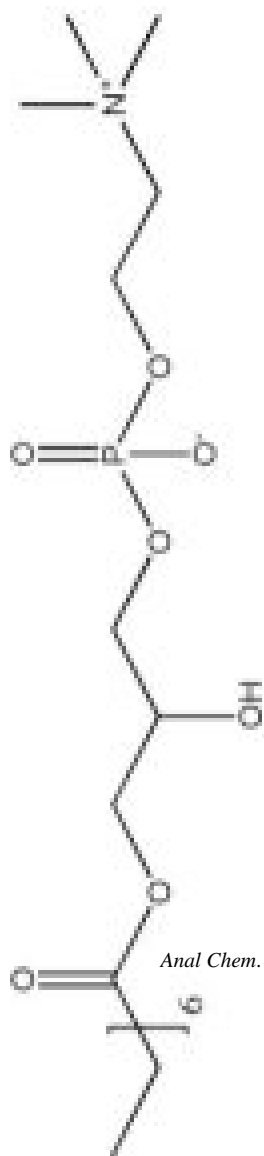
Related molecule(mass)

Identified common structure

Glycerophosphocholine (257 Da)

Fragment score = 54   Neutral loss score = 5

returned due to the database limitations.
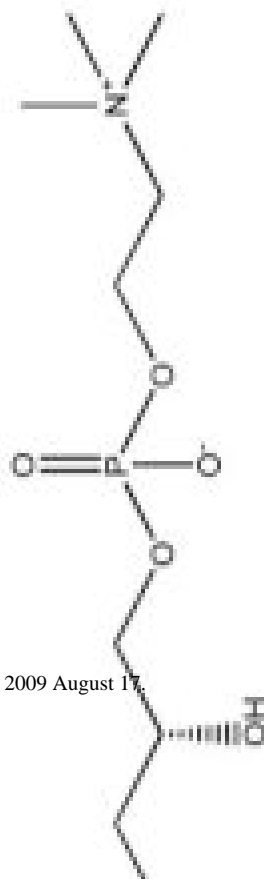
Related Group

lecule(mass)

Related Group

ecule(mass)

1-Palmitoyllysophosphatidylcholine (495)
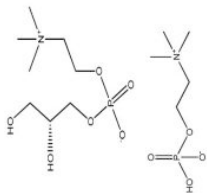
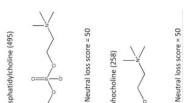Fragment score =93 Neutral loss score = 8

Glycerophosphocholine (258)

Fragment score =79 Neutral loss score = 25

Related Group

ecule (mass)

[a] From the results of substructure and searching the accurate mass, two possible metabolites were identified (509 m/z and 610 m/z).