METLIN Personal Compound Database and Library (PCDL) for Metabolite Identification Using Accurate Mass, Retention Time, and MS/MS Spectra Matching

Technical Overview

Authors

Steve Fischer
Agilent Technologies, Inc.
Santa Clara, CA USA

Theodore Sana
Agilent Technologies, Inc.
Santa Clara, CA USA

Introduction

A key step in metabolomics research continues to be the identification of unknown compounds. The level of confidence in the identification is directly dependent on the amount of analytical information that can be used to uniquely identify an unknown. When collecting LC/MS data there are four pieces of information available to identify unknown compounds: monoisotopic mass, isotopic mass values and abundance ratios, chromatographic retention time, and MS/MS fragmentation spectra. Historically, researchers looking to identify unknown compounds have used accurate-mass measurements (<5 ppm) of unknowns to search a database of metabolites. While this excludes a large number of metabolite possibilities, the match is by no means sufficient to identify unknown metabolites. Including isotopic-mass values and abundance ratio measurements increases the specificity of the provisional metabolite database search result. However, this only establishes the empirical formula of the unknown compound and a list of possible metabolites that are consistent with that empirical formula. Additional orthogonal information, such as retention time and MS/MS spectra, can be used to increase confidence in compound identification. This technical overview illustrates the use of accurate-mass retention time and MS/MS spectra matching with the METLIN PCDL in order to identify metabolites with much higher confidence (Figure 1).

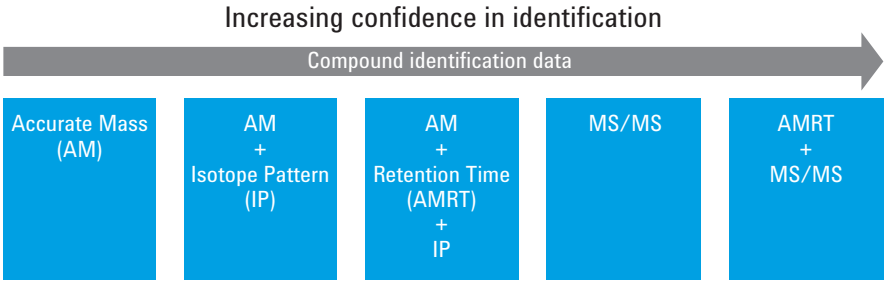

Figure 1. The addition of retention time and accurate-mass MS/MS spectra matching information increases confidence in compound identification.

Agilent Technologies

The METLIN Metabolite Personal Compound Database and Library

Agilent Technologies is the exclusive provider of the METLIN Personal Compound Database and Library (PCDL), which is part of a continuing collaboration between Agilent Technologies and Dr. Gary Siuzdak. The METLIN PCDL resides on your PC, working with other Agilent software packages to facilitate fast and easy compound searching (manual or automated). The METLIN PCDL was created by adding MS/MS spectra to the METLIN Personal Compound Database (PCD), which contains approximately 25,000 compounds, of which more than 670 include measured retention times based on chemical standards. The METLIN PCDL also contains more than 2270 compounds with accurate-mass quadrupole time-of-flight (Q-TOF) MS/MS reference spectra to provide the highest confidence in metabolite identifications. Unlike unmonitored Web-based databases, with open source MS/MS spectra, the METLIN PCDL has been subjected to rigorous quality control to ensure metabolite identifications that you can trust.

Quality data for confidence in identification

The correct identification of unknown compounds requires accurate LC/MS measurements, and a database, or library, with sufficient size and high-quality reference data. Great care has been taken to include a large number of relevant compounds in the PCDL along with empirical formulae, reproducible retention times, and accurate MS/MS spectra. Retention times were generated on an Agilent 1200SL system with chromatographic conditions suitable for a wide variety of metabolites and validated on multiple LC/MS systems and columns. The MS/MS spectra were generated from standards, using multiple Agilent Q-TOF LC/MS instruments and with a set of conditions that ensures comprehensive metabolite coverage. All data was subjected to rigorous quality control before addition to the PCDL, resulting in a high-quality database and library that enables higher confidence in metabolite identifications than Web-based searches.

Adding retention times to the METLIN PCD

The retention times for the standards have been carefully determined using uniform chromatographic conditions (Appendix Table 1). A simple, linear reverse-phase gradient that is applicable to the vast majority of metabolite compounds was used to acquire retention time data. The conditions were designed for use with electrospray (ESI) and atmospheric chemical ionization (APCI), as well as positive and negative mode ionization (Appendix Table 2). To verify usability and quality, three different LC/MS systems and three different column lots were used for data collection, and the extracted ion chromatogram (EIC) of replicate injections of each standard was evaluated by a chemist. The mass and isotopic pattern were then confirmed and the average retention time was added to the metabolite listed in the METLIN PCD. This level of curation ensures a quality-controlled database that delivers higher confidence in compound identification.

Adding MS/MS spectra to the METLIN PCDL

While accurate-mass and retention-time matching alone can provide high-confidence identification for most compounds, adding the MS/MS library searching capability provides a chromatography-independent means of compound identification. The METLIN PCDL contains accurate-mass MS/MS reference spectra for more than 2270 standards. When this information is combined with accurate-mass molecular ion measurements and chromatographic retention time, the METLIN PCDL provides the highest confidence in metabolite identifications.

Multiple collision energies for comprehensive metabolite coverage

A comprehensive approach was used to produce an MS/MS spectral library capable of handling the challenges of metabolomics. The library entries were generated from standards using an ESI source on multiple Agilent Q-TOF LC/MS instruments to ensure reproducible MS/MS reference spectra. In addition, since most metabolites ionize strongly in one polarity and poorly in the other, spectra were collected in both positive and negative ion modes. Since some compounds fragment easily, while others require more energy to fragment efficiently, spectra were also collected at three collision energies: 10, 20, and 40 eV. This thorough data collection strategy ensures useful spectral data coverage for a wide range of metabolites with different physico-chemical properties. For example, stearic acid is a long-chain fatty acid that requires 40 eV to fragment, whereas hippuric acid fragments easily at 10 eV (Figure 2). In addition, compounds such as glutamic acid ionize in both positive and negative ion modes, but produce very different MS/MS spectra (Figure 3).

For analysis of the standards, the MS/MS spectra were all produced from the isolated monoisotopic ion by setting the quadrupole filter of the Agilent 6520 Q-TOF LC/MS to high resolution (peak width = 1.3 amu), so that the adjacent, naturally occurring isotopes were excluded. This increases the specificity of the MS/MS spectral library. The instrument conditions used to generate the MS/MS spectra are shown in Appendix Table 3.

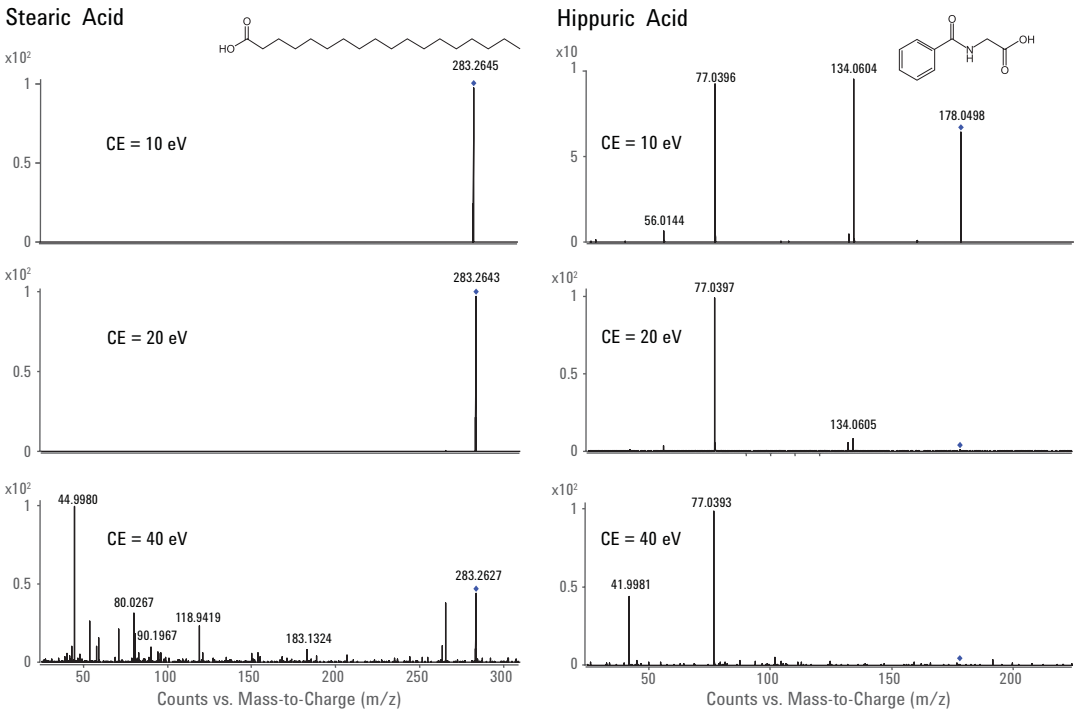

Figure 2. MS/MS spectra for stearic and hippuric acid illustrate the use of multiple collision energies (10, 20, and 40 eV) to assure complete spectral data coverage.

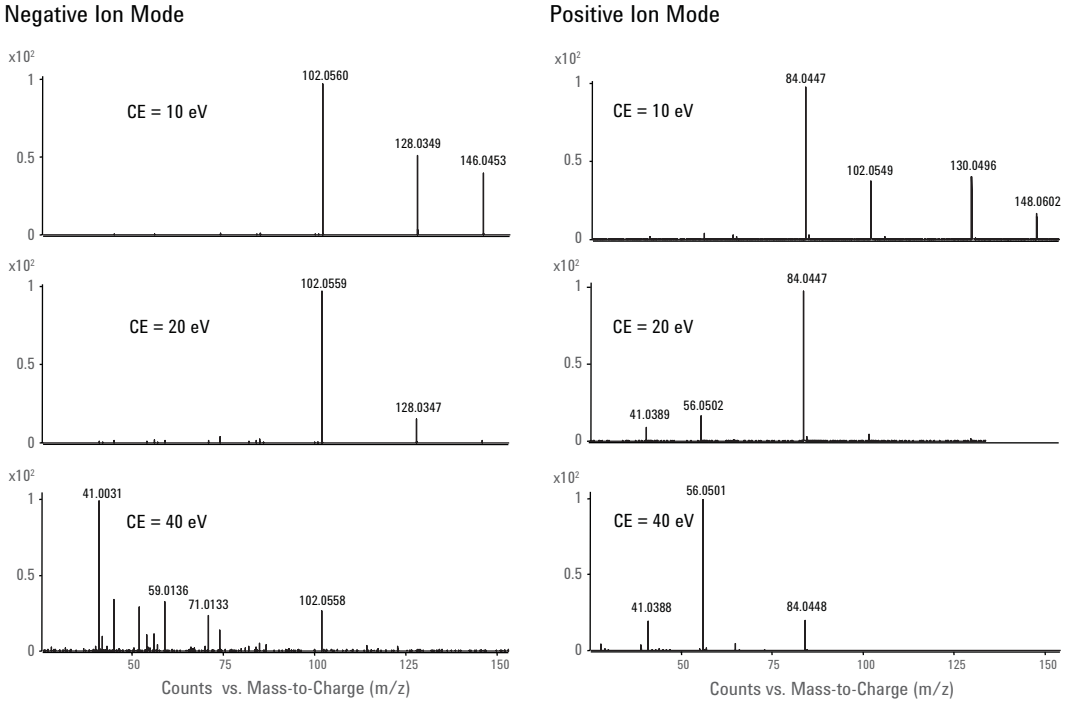

Figure 3. MS/MS spectra for glutamic acid generated using multiple collision energies (10, 20, and 40 eV) to ensure useful spectral data coverage in both positive and negative ion modes.

Quality control of MS/MS spectra

Prior to their addition to the METLIN PCDL, the MS/MS spectra generated for the standards were quality checked using a proprietary algorithm. The spectra were filtered using an absolute threshold of 100 counts and a relative threshold of 1.0 % of the largest ion in the spectrum. This process filtered out the weak ions that were either insignificant or too weak for robust library matching. The remaining ions were then tested for agreement with the empirical formula of the known compound. Only those ions that passed these quality control tests were considered to be related to the precursor ion and retained. The final library spectrum contains the mass and intensity of each ion with the fragment masses corrected to their theoretical mass value.

Every original acquired spectrum was searched against the library spectrum using forward and reverse searching to assure that the acquired spectrum matched the imported spectrum. The search results were then manually reviewed to determine if the library spectrum was of high match quality or should be excluded due to poor match quality. This process resulted in the retention of only those spectra that were of very high quality. In addition, many compounds have fewer than the six expected library spectra (three collision energies x two ion modes) because spectra were excluded if they did not contain information that was considered useable.

Using the METLIN PCDL

Identification of metabolites in human urine using AMRT matching

As an orthogonal measurement to accurate mass, retention time serves to increase the confidence with which PCDL database matches are made. The total ion chromatogram (TIC) of a urine sample was shown to be a complex chromatogram containing many peaks (Figure 4A). The extracted ion chromatogram (EIC) of m/z 206.0827 revealed four distinct peaks with different retention times that are likely to be structural isomers with the same empirical formula, C11H11NO3 (Figure 4B). This example demonstrates that accurate-mass matching alone is not sufficiently discriminating for confident compound identification.

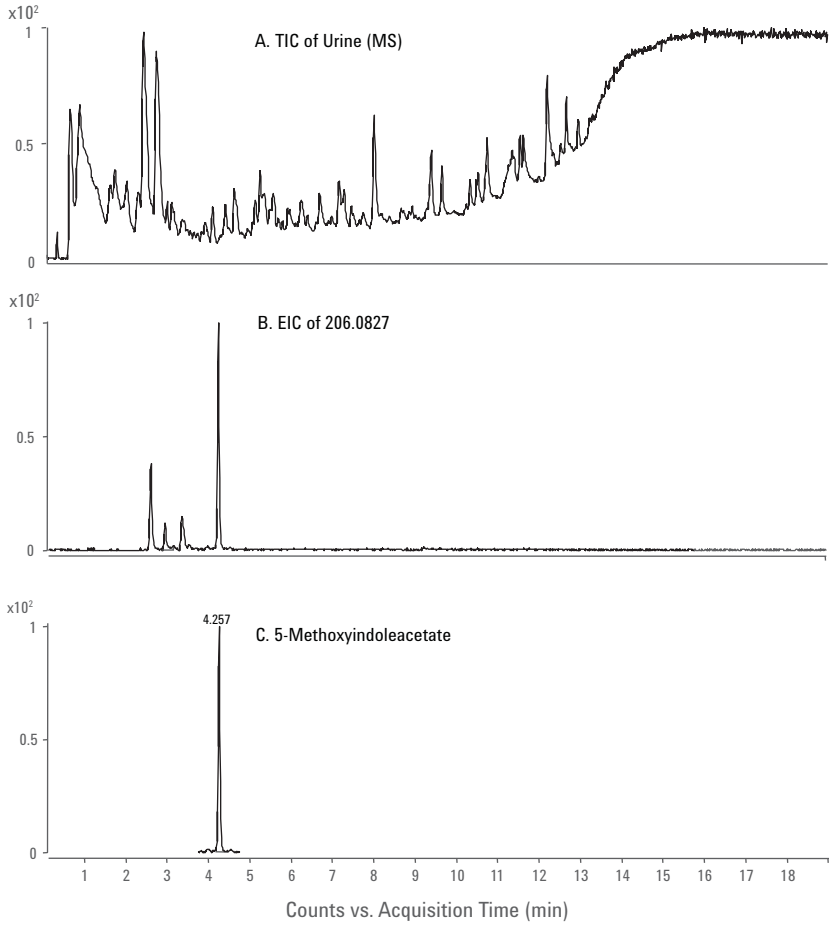

Figure 4. Using accurate-mass and retention time (AMRT) matching to analyze metabolites in urine. (A) Total ion chromatogram (TIC). (B) Extracted ion chromatogram (EIC) of accurate mass 206.0827, showing 4 peaks. (C) Matching to the retention time of 5-methoxyindoleacetate (4.257 min) resulted in a positive identification.

The addition of compound retention time provides an additional parameter that, in most cases, is sufficient to distinguish a compound from others of the same mass. When the METLIN accurate-mass retention time (AMRT) database was searched again, requiring both accurate-mass and retention time matching, only a single peak satisfied both parameters. The formula $C_{11}H_{11}NO_3$ was annotated from the database as 5-methoxyindoleacetate (Figure 4C). This illustrates the powerful ability of AMRT matching to quickly find a high-confidence match to the database by filtering away isomers and other compounds with the same empirical formula that do not satisfy the retention time requirement.

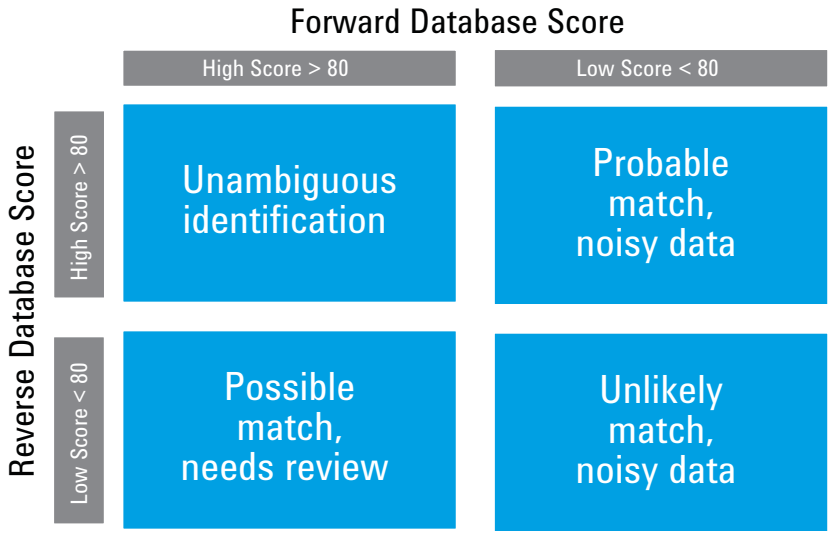

Figure 5. Confidence of compound identification based on reverse and forward library match scoring.

Identification of metabolites using MS/MS library matching

Spectral matching for each compound involved selecting only those compounds in the library that matched the mass of the precursor ion and then performing a mathematical comparison. A dot product score is calculated by comparing the library and unknown peak intensities for all of the ions found in the observed spectrum within the mass tolerance that matches to the library spectrum. The result is a match quality score of 0-100, with 100 being a perfect match.

MS/MS library matching is done by comparing the library to the data (reverse search) or the data to the library (forward search). Reverse searching tests for the presence of the library ions in the observed spectrum. It is used to detect the presence of a compound in a MS/MS spectrum. Forward searching tests for the presence of the spectrum ions in the MS/MS library spectrum. It is used to highlight a noisy spectrum that may imply the observed spectrum is impure and the reverse search match is suspect. For example, a reverse score of 80 or greater indicates a match and a forward score of 50 or greater indicates a relatively pure MS/MS spectrum.

Metabolite identification in erythrocyte and arabidopsis extracts

Two different complex samples, malaria-infected red blood cell extract and whole leaf arabidopsis extract, were processed and analyzed by LC/MS and LC/MS/MS to illustrate the ability of the METLIN PCDL to identify compounds by MS/MS spectral matching. The chromatography and Q-TOF LC/MS conditions, and parameters used for these analyses, were the same as those shown in Appendix Tables 1-3. The matching of compound spectra obtained from the samples to library spectra of the standards was performed in MassHunter Qualitative Analysis software B.04, using both forward search (match all ions in unknown spectra to the library) and reverse search (match only library ions to the spectra of the unknown) approaches (Figure 5).

Targeted MS/MS analysis was performed on a number of ions eluting between 0.799 and 1.330 minutes found in an erythrocyte extract minutes. The search results for one of the compounds (m/z = 130.0499, signal 14,000 counts) analyzed using targeted MS/MS (Figure 6) gives a high confidence assignment of pyroglutamic acid (CE = 20 eV number of ions = 7, reverse score = 89, and forward score = 61).

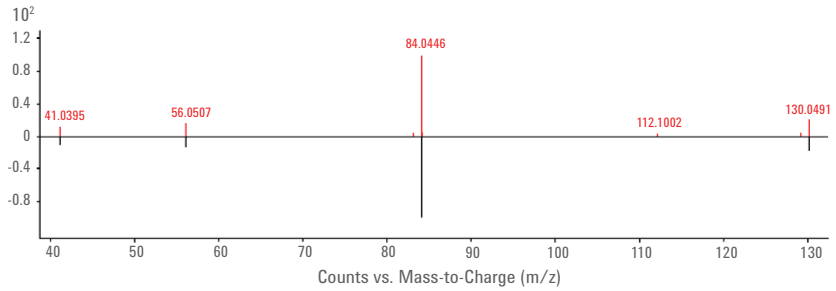

Figure 6. Mirror image display of the library match results for the 20 eV MS/MS spectrum of m/z 130.0499 observed in an erythrocyte extract with a retention time of 0.97 min. The ion was identified as pyroglutamic acid.

Targeted MS/MS analysis was also performed on a compound in the arabidopsis extract (m/z = 611.1629, signal 108,000 counts) with a retention time of 5.18 min. The search results for the compound analyzed using targeted MS/MS (Figure 7) gives a high confidence assignment of rutin (CE = 20 eV, number of ions = 6, reverse score = 99, and forward score = 94). The scores for rutin are higher than for the previous example, pyroglutamic acid, because of the stronger MS precursor signal available for fragmentation (108,000 versus 14,000 counts).

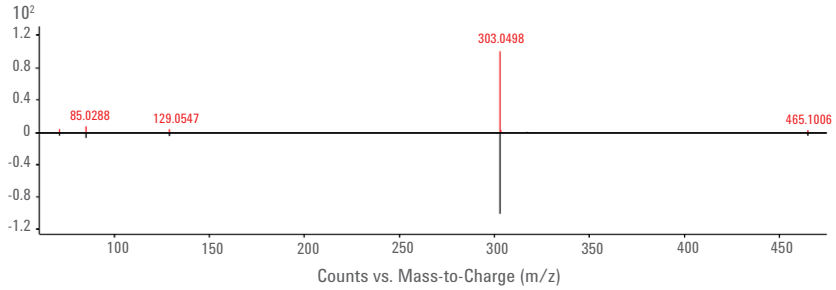

Figure 7. Mirror image display of the library match results for the 20 eV MS/MS spectrum of m/z 611.1629 observed in an arabidopsis extract with a retention time of 5.18 min. The ion was identified as rutin.

Conclusions

Accurate identification of unknown compounds requires a large, high-quality database that is well curated. The Agilent METLIN Personal Compound Database (PCD), which contains more than 25,000 metabolites supplemented with the retention times of more than 670 standards, provides personalized access to the largest metabolomics-focused database. The Agilent METLIN PCDL supplements this with more than 2270 MS/MS spectra to greatly increase the accuracy of identifications. Rigorous quality control procedures have been instituted throughout in order to ensure the accuracy of the database and library contents. Chromatographic conditions suitable for a wide variety of metabolites were used to determine retention times. To validate the retention times, multiple Agilent Q-TOF LC/MS systems and column lots were used. The MS/MS spectra were generated using three collision energies, in both positive and negative ion modes, to ensure comprehensive metabolite coverage. Attention to quality provides assurance that the Agilent Personal Compound Database and Library will yield identifications with much higher confidence than those obtained from Web-based resources.

Appendix

Table 1. LC analysis conditions used for determining retention times of standards.

LC Instrumentation	
LC system	Agilent 1200SL Series HPLC system
Stainless steel connecting capillary	700 mm, 0.17 mm (green) instead as outlet capillary or pump to injector, p/n: G1312-87304
Capillary connecting autosampler to column compartment	340 mm, 0.12 mm id (red), p/n: G1316-87319 directly to the heater block in the heated column compartment or 300 mm, 0.12 mm id (red), p/n: G1316-87318 directly to the switching valve in the heated column compartment
Heater to guard column and guard column to analytical column connections	Stainless steel tubing, 70 mm X 0.12 mm ID (red label), p/n: G1316-87303
Analytical column to MS connection	PEEK Tubing (red), 0.005"/0.13 mm, 1/16 OD, 5.0 m, cut to 650 mm in length, p/n: 5042-6461
Guard column	Zorbax-SB-C8 Rapid Resolution Cartridge, 2.1 X 30 mm, 3.5 µm, p/n: 873700-936
Column	Zorbax SB-Aq, 1.8 µm 2.1 X 50 mm, p/n: 827700-914
LC Conditions	
Column temperature	60 °C
Autosampler temperature	4 °C
Injection volume	5 µL
Mobile phase	A = H_2O + 0.2 % acetic acid B = methanol + 0.2 % acetic acid
Flow rate	0.6 mL/min
Gradient	0 – 13 min, 2.0 % – 98 % B 13 – 19 min, 98 % B hold
Stop time	19 min
Post time	5 min

Additional Chromatography Instructions:

Samples were dissolved in 50:50 methanol:water, 0.2 % acetic acid. Methanol was added to the sample first and then vortexed vigorously. Next, water containing 0.4 % acetic acid was added to get to the final composition. This method usually covers most metabolite compounds, except for very hydrophobic molecules.

(2) The mixer and pulse damper was used in bypass mode as described in the Agilent 1200 Series Binary Pump SL User Manual (p/n: G1312-90011), page 75.

(3) A minimal delay volume was used in the 1200SL system.

(4) The left and right side of the heater on the column compartment were set to the same temperature, the column was pressed tightly against the heater block with metal clips, and the column compartment cover was put in place.

Table 2. MS analysis conditions used with retention time chromatography to determine accurate-mass of metabolite standards.

MS Conditions	
Instrument	Agilent 6520 Accurate-Mass Q-TOF
Ion mode	Dual ESI source in both positive and negative ion mode
Dynamic mass axis calibration	Continuous infusion of a reference mass solution using an isocratic pump connected to a dual sprayer electrospray ionization source
Reference ions	Negative ion mode: 119.0363, 980.016375 Positive ion mode: 121.0509, 922.0098
Drying gas temperature	325 °C
Drying gas flow	10 L/min
Nebulizer pressure	45 psi
Vcap	4000 V (positive), 3500 V (negative)
Fragmentor voltage	140 V
Mass range	20 – 1650 amu
Acquisition rate	2 Hz

Table 3. MS analysis conditions used for determining MS/MS spectra of standard compounds.

MS Conditions	
Instrument	Agilent 6520 Accurate-Mass Q-TOF
Ion mode	Dual ESI source in both positive and negative ion mode
Dynamic mass axis calibration	Continuous infusion of a reference mass solution using an isocratic pump connected to a dual sprayer electrospray ionization source
Reference ions	Negative ion mode: 119.0363, 980.016375 Positive ion mode: 121.0509, 922.0098
Drying gas temperature	325 °C
Drying gas flow	10 L/min
Nebulizer pressure	45 psi
Vcap	4000 V (positive), 3500 V (negative)
Fragmentor voltage	140 V
Mass range	20 – 1650 amu
Acquisition rate	2 Hz

www.agilent.com/chem/metlin

Published in USA, August 24, 2011
5990-8918EN

Agilent Technologies