Applying In-Silico Retention Index and Mass Spectra Matching for Identification of Unknown Metabolites in Accurate Mass GC-TOF Mass Spectrometry

ARTICLE *in* ANALYTICAL CHEMISTRY · JUNE 2011

Impact Factor: 5.64 · DOI: 10.1021/ac2006137 · Source: PubMed

Doug Stevens
Waters Corporation
5 PUBLICATIONS **37** CITATIONS
SEE PROFILE

Tobias Kind
University of California, Davis
108 PUBLICATIONS **2,537** CITATIONS
SEE PROFILE

Carsten Denkert
Charité Universitätsmedizin Berlin
308 PUBLICATIONS **8,531** CITATIONS
SEE PROFILE

Oliver Fiehn
University of California, Davis
293 PUBLICATIONS **17,482** CITATIONS
SEE PROFILE

Published in final edited form as:
Anal Chem. 2011 August 1; 83(15): 5895–5902. doi:10.1021/ac2006137.

Applying *in-silico* retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry

Sangeeta Kumari[1], **Doug Stevens**[2], **Tobias Kind**[1], **Carsten Denkert**[3], and **Oliver Fiehn**[1]

[1]UC Davis Genome Center, Davis, CA 95616, USA

[2]Waters Corp. MA

[3]Charite University Clinics, Berlin, Germany

Abstract

One of the major obstacles in metabolomics is the identification of unknown metabolites. We tested constraints for re-identifying the correct structures of 29 known metabolite peaks from GCT premier accurate mass chemical ionization GC-TOF mass spectrometry data without any use of mass spectral libraries. Correct elemental formulas were retrieved within the top-3 hits for most molecular ion adducts using the "Seven Golden Rules" algorithm. An average of 514 potential structures per formula was downloaded from the PubChem chemical database and *in-silico* derivatized using the ChemAxon software package. After chemical curation, Kovats retention indices (RI) were predicted for up to 747 potential structures per formula using the NIST MS group contribution algorithm and corrected for contribution of trimethylsilyl groups using the Fiehnlib RI library. When matching the range of predicted RI values against the experimentally determined peak retention, all but three incorrect formulas were excluded. For all remaining isomeric structures, accurate mass electron ionization spectra were predicted using the MassFrontier software and scored against experimental spectra. Using a mass error window of 10 ppm for fragment ions, 89% of all isomeric structures were removed and the correct structure was reported in 73% within the top-5 hits of the cases.

Introduction

Identification of unknown metabolites is the major bottleneck in the field of metabolomics[1–4]. Most of the detected metabolites are unidentified due to small libraries and unavailability of reference standards[5]. The current effort to solve this problem involves creation of extensive libraries[6–8], use of different instrument platforms (LC-NMR[9,10], LC-TOFMS[11,12], GCTOF-MS[13], CE-TOFMS[14,15]), development of algorithms for calculation of retention indices[16] and software for prediction of *in-silico* spectra for compounds[17,18].

Metabolomic studies often use GC-MS techniques because of wide spread availability of instruments, broad coverage of metabolites comprising most primary metabolites (except di- and triphosphates), ease of use (when following high quality standard operating procedures (SOP)[19], good reproducibility of quantitative results and because of the availability of extensive mass spectral libraries[20]. Nevertheless, GC-MS based metabolomics requires

Corresponding Author: ofiehn@ucdavis.edu.

Competing interest's statement

The authors declare no competing financial interests.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

derivatization for all compounds that are not readily volatile[21], most commonly using trimethylsilylation (TMS) of acidic protons to exclude hydrogen bonding and hence lower boiling points[21]. The reactivity of different functional groups towards trimethylsilylation varies in the following order: alcohol (primary> secondary> tertiary)> phenols> carboxylic acids> terminal amines> acid amides> secondary amines[22]. Specifically for amino groups, TMS can replace two acidic hydrogens, only the most acidic one or sometimes even none of them.

Identification of unknown metabolites has to start from obtaining elemental composition formulas. We have previously reported that chemical ionization accurate mass GC-TOF mass spectrometry yields the correct formula at high likelihood within the first 3 hits using a constraint-based algorithm, specifically using isotope ratio data that complement accurate mass information[23,24]. Further approaches have been proposed using prediction of retention indices[25–27] or *in-silico* generation of mass spectra for compounds[28]. Neither of these two approaches are very straightforward to be implemented in GC-MS based metabolomic assays because of the TMS-derivatization which dominates behavior for both chromatographic retention and mass spectral fragmentation[29].

Nevertheless, important procedures have already been validated. For example, it has been shown that retention indices that are relying on different marker compounds can be easily converted within a small error range, for example from linear aliphates (Kovats) to fatty acid methyl esters or even polyaromatic hydrocarbons[30]. The retention behavior of any compound depends on the molecular interaction between the solute and stationary phase[31] and rely on the chemical structural properties of the compounds[32–34]. A combination of retention index prediction with further classifiers in electron ionization GC-MS enabled drastic reduction of all possible candidate structures for $C_{12}H_{10}O_2$[29]. Hence, it is feasible and useful to calculate predicted retention indices for a given compound from their molecular features[27,35,36]. Among these RI calculation approaches, the ‘group contribution method’[37] has been made freely available to use within the NIST MS software which was built based upon experimental retention indices for over 35,000 different molecules. As molecular structure features impact retention in many different ways, no RI prediction software to date can correctly predict the exact retention difference for isomeric structures. Therefore, we here show how lists of structural isomers can be further constrained using mass spectral fragmentation predictions.

Mass spectra in GC-MS are very difficult to predict, specifically when using hard ionization techniques such as the most commonly applied electron ionization at 70 eV. Molecules fragment in a range of routes, and more importantly, fragments can also follow rearrangement reactions. Approaches have extended to fragmenting every single bond in molecular structures[28] or the accumulation and application of all published fragmentation pathways[18] which has been made available as Mass Frontier software package. In GC-MS, application of Mass Frontier has been very rarely used so far, likely due to the complexities of derivatizations and rearrangements of fragment radicals. We here show that theoretical fragmentations can still be useful to constrain hit lists of isomeric structures if accurate mass data can be used that would limit scoring results in matching predicted and experimental mass spectra.

Experimental Methods

A mixture of 1mg each of reference compounds was prepared in 1 ml of water/methanol/ isopropanol as solvent, comprising compounds given in table 1. 10 microliters of this mixture were dried down and prepared for analysis in a two-reaction derivatization schema [19] for gas chromatography/orthogonal time-of-flight mass spectrometry (Waters

Micromass GCT Premier, Milford, MA, USA) under electron ionization and ammonia, isobutane and methane chemical ionization as published previously (supplement 1).

For each GC-MS peak, accurate masses of the molecular ion adducts were determined which were found as $[M+H]^+$ when using ammonia, methane or isobutane as chemical reagents in chemical ionization. After obtaining the top-3 hits for potential elemental formulas, all corresponding structures were downloaded using the 'molecular formula' search from the PubChem database compound identifiers (CID)[38] as 'structure data format' files (SDF). The ChemAxon Standardizer tool[39] was then applied to curate these SDF structures by removing structure fragments, isotope atoms, any attached data and explicit hydrogen bonds. Aromatic structures were formalized with explicit double bonds. The resulting standardized structures were first methoximated using the Standardizer tool[39] and then trimethylsilylated in a step-wise manner using the ChemAxon Reactor tool[5,40]. All derivatized structures obtained in this manner were imported into the ChemAxon Instant JChem tool[41] to perform substructure searches: all isomers that still had free hydroxyl and carboxyl groups were removed because the reactivity of MSTFA (N-methyl-N-(trimethylsilyl)trifluoroacetamide) is known to completely derivatize such groups, whereas amines were retained as fully, partly or non-derivatized structures. All structures were removed that had lower or higher molecular masses than experimentally observed. These curated structures were used to calculate predicted retention index (RI) values using the NIST RI algorithm with a correction factor that was obtained as follows: 285 structures of the Leco Fiehnlib library[42] were downloaded from PubChem and subjected to *in-silico* derivatization as described above using the Standardizer tool[39]. The derivatized structures were then imported into the Instant JChem tool[41] and manually curated for the correct number of trimethylsilyl groups. The curated structures were exported as SDF file from Instant JChem and subjected to RI calculation using the NIST RI algorithm. The differences between experimental and predicted RI values were plotted against the number of trimethylsilyl groups for determining an additional group contribution correction.

All structures remaining after retention index filtering were subjected as SDF files to the *in-silico* fragmentation in Mass Frontier in batch mode. Fragmentations were performed using general fragmentation rules allowing up to five secondary reactions to obtain rich fragmentation spectra. *In-silico* spectra were scored against accurate mass electron ionization experimental spectra using a 10 ppm mass accuracy threshold.

Mass Frontier Software: Mass Frontier 6.0 software [HighChem, Ltd. Slovakia, (vs. 6.0.2.5) 2009] has been used for prediction of fragmentation pathways. Generation of fragments for all candidate structures using Mass Frontier was implemented using the Batch Processing function, SDF file containing all candidate structures as input. As knowledge base, general fragmentation rules have been applied. The following settings options (under reaction restriction option) were used: Knowledge Base Tab: General fragmentation rules only. Ionization and cleavage Tab: Default settings. H-Rearrangement tab: Default settings. Resonance Tab: Default settings. Additional Tab: Default settings. Size Tab: Reaction Steps: 5, Mass Range: 30 to 1000m/z, Reaction Limits: value 10000. The output option 'add to database manager' is selected with database manager option 'Annotate spectra'.

Result and discussion

Over 160 metabolites in breast cancer tissues can be directly identified by GC-TOF MS using electron ionization spectral matching

Untargeted metabolomic studies yield hundreds of unknown peaks. It is important to first identify well-known metabolites such as amino acids, sugars, lipids and hydroxyl acids[43–45] by retention-based spectral libraries before using more advanced techniques. Often, the number of unknowns in published reports exceeds the number of identified metabolites by a factor of 2–3 even after extensive data processing[44]. As example, we have screened breast cancer tissues for differences in cancer grades and in comparison to healthy tissue (supplement 2). Using our BinBase database[19], 161 unique identified metabolites plus additional 239 unknown metabolite peaks were detected in low resolution GC-TOF MS screens (after exclusion of known artifacts such as phthalates or polysiloxanes). BinBase is supported by the NIST08 MS and Fiehnlib[42] libraries for identifying compounds based on retention index and mass spectral matching. An additional query against the MassBank database[46] resulted in the annotation of the previously unidentified peaks BinBase #227600 as uridinediphosphate-N-acetylglucosamine (UDP-GlcNAc) and BinBase #328006 as UDP-glucuronate which were confirmed by comparison to commercial reference standards. UDP-GlcNAc is an important substrate for formation of protein glycans and has been repeatedly highlighted as involved in cancer progression[47]. The identification of UDP-GlcNAc serves as example that it is important that all available libraries are screened as first-screen annotations. Secondly, UDP-GlcNAc and UDP-glucuronate shows that mass spectral libraries (like MassBank) may annotate peaks by their authentic structures while in fact, the observed peaks are thermolytic breakdown products as evidenced by CI-based GC-MS (see supplement 2). Most bisphosphates and other labile metabolites do not withstand the hot conditions in GC-MS operation[43]. Thirdly, electron ionization spectra often do not yield clearly identifiable molecular ions and thus, soft ionization techniques must be employed for structure dereplication of unknown peaks.

Impact of reagent gases on ion abundance and mass accuracy

We therefore used a test mixture of known metabolites to explore how well structural annotations of unknown peaks in GC-MS chromatograms could perform with current instrumentation and tools but without any further *a priori* information such as spectral libraries. Chemical ionization yield molecular proton adducts. Depending on the relative proton affinity of the metabolites and the reagent gas ions[48], overall ionization efficiency differs depending on chemical classes. In metabolomics, many different structures are investigated. We have thus used a representative mixture of structurally diverse metabolites to test ionization efficiencies (and mass accuracy) when using different reagent gases (supplement 3) in the Waters GCT premier TOF-MS instrument with the most commonly used reagent gases, namely ammonia, isobutane and methane. As expected, molecular ions were mostly not detectable or low abundant in EI.

Under CI conditions, a total of 29 metabolite peaks were used to determine mass and isotope accuracy under CI ionization (supplement 3). A third of all molecular ion adducts were found at abundant intensity values (>11,000 counts per averaged spectrum, cps), 40% at moderate intensities (>1000 cps, <11,000 cps) and 26% at low abundance of <1,000 cps. Indeed, 26 of the 29 molecular adducts were more than 2-fold lower abundant under methane CI compared to isobutane CI, and no compound showed more than 2-fold higher abundance. The median abundance was 97% comparing ammonia CI to isobutane CI. From this data we conclude that ammonia and isobutane can be recommended as CI reagent gases for trimethylsilylated metabolite analysis, but not methane. Mass accuracy and isotopic abundance accuracy for all molecular ion peaks were not dependent on the reagent gas, confirming data published before[24] with an overall mass error of 2.3±2.8 mDa and an isotopic abundance ratio error for A+1/A and A+2/A as 2.0±2.6 % and 1.5±1.7 %, respectively (average and standard deviations, supplement 3).

The Seven Golden Rules algorithm yields the correct elemental formula within the top-3 hits

The mass accuracy available with the instrument used here is not sufficient even at masses below 500 Da to obtain the correct molecular formula as top hit[49]. Elemental formula calculators such as MWTWIN[50] yield up to 1,000 elemental compositions for the compounds we have used here, depending on element search boundaries. We have therefore integrated a range of constraints into the calculation of elemental formulas from accurate mass analyses such as isotope ratio abundance, chemical rules and heuristic atom ratios obtained and made the software freely available as Seven Golden Rules algorithm[23]. In order to include 95% of all peaks, one needs to use two standard deviations as search window boundaries for the instrument error estimates; otherwise, one third of all true hits would be excluded. Despite using a very wide search range (20 ppm mass errors with 10% isotope ratio error), we have obtained the correct formula within the top-3 hits in 27/29 test cases using the Seven Golden Rules algorithm[23]. Indeed, 22/29 of our test metabolite peaks were retrieved as top-hit. Although a very high number of formulas were retrieved as first-hit, for unknowns one cannot assume that this will be the case, especially because unknown metabolite peaks tend to be of lower abundance and hence will have higher errors in data acquisition. Therefore, it is advisable to query at least the first three elemental compositions for potential structure matches.

Hundreds of structures are obtained from chemical database queries

For identifying unknowns, the next step is to search for potential structures that might represent the observed metabolite peaks. We have outlined the overall workflow from 'accurate mass spectrum' to 'best structure hit' in figure 1. Our test case metabolites are all contained in the comparatively small biochemical databases MetaCyc, HMDB and KEGG, but for most unknowns we need to assume that these might represent very rare small molecules. Hence, larger chemical databases need to be queried such as the public PubChem repository. Almost all molecules in PubChem are presented only in their non-derivatized form. Therefore, queries first need to remove the derivatization groups from the elemental formulas obtained by the Seven Golden Rules algorithm. The trimethylsilyl groups can be easily removed by subtracting SiC_3H_8 for each silicon atom in the elemental formula. In GC-MS based metabolomics, keto- and aldehyde groups, e.g. in sugars, are protected by methoximation in order to enable chromatographic separation of isomeric structures[51]. Unknown peaks that bear nitrogen in their formulas therefore need to be tested chemically if the nitrogen stems from the methoxime derivatization, using ethoxyamine instead of methoxyamine in a replicate biological sample. Ethoximation causes derivatized carbonyl-containing metabolites to shift towards higher retention times and also causes shifts in mass spectra[44]. Hence, the use of derivatization reactions enables recognizing chemical groups in unknown structures (i.e. the number of acidic protons and carbonyl groups). This information can be used as constraint in the workflow, unlike for underivatized molecules as they are common in LC-MS based metabolomics.

The non-derivatized elemental formulas were then queried against PubChem using the 'molecular formula' search within the tool. On average, 508 structures were derived and downloaded as SDF files for the 87 formulas we have investigated (table 1, full content as supplement 4). For example, 570 structures were retrieved matching the elemental composition of lysine ($C_6H_{14}N_2O_2$). However, many of these structures could be excluded from further searches as PubChem also contains structures with additional fragments (like water molecules) or natural isotopes (e.g. deuterium and tritium are listed as hydrogens in the formulas). Examples are given in figure 2. These structures are removed using the 'Standardizer' tool[39] (fig. 1, step 2) which is implemented in ChemAxon's JChem software[41]. In addition, we standardized all structures by neutralizing charged structures, writing explicit double bonds for aromatic structures and replacing implicit hydrogens by adding these as explicit atoms in the structure files. All stereocenters were replaced by the corresponding achiral structures as the gas chromatography used in our test cases would not resolve chiral isomers. All aldehyde- and keto-groups that did not have adjacent heteroatoms were standardized as methoxime groups. Next, all structures were *in-silico* derivatized using ChemAxon's Reactor tool[40] (step 3 in figure 1) by replacing all H-donor moieties with trimethylsilyl groups in a step-wise manner.

Afterwards, structures were curated (step 4, figure 1) with respect to (a) the known chemistry of the trimethylsilylation and (b) concurrence of the observed accurate mass with the mass of the *in-silico* derivatized structures. During this curation step, all structures were removed that had underivatized hydroxyl-, thiol- or carboxyl-groups because the trimethylsilylation reaction is known to perform exhaustively and complete for these groups (fig. 2). In contrary, amines are known to yield peaks that can be left underivatized or be replaced with one or two trimethylsilyl groups. While the peak area ratio of partly and fully derivatized amines can be controlled by rigorous quality control of the GC-MS injection system as published before[19], such precautions will not lead to the detection of only one peak per amine for most metabolites. Consequently, the curation step left all amines in place independent of their trimethylsilylation status.

Subsequently, derivatized structures were removed that yielded different accurate masses than the experimentally observed mass. During the structure standardization, derivatization and curation steps, on average 79% of all potential PubChem structures per formula were successfully removed by not matching the structure constraints (e.g. presence of functional groups). Creatinine 3TMS was not found as candidate structure after applying the Standardizer and Reactor tools because the PubChem structure for creatinine lists one acidic proton, while the actual trimethylsilylation results from three H-donor moieties due to tautomerization of the parent structure (supplement 5). At this point, a correct consideration of tautomers is beyond the capability of the Standardizer tool.

Retention index prediction constrains hit lists

Steps 1–4 in the identification workflow (fig. 1) have yielded up to 747 derivatized structures per elemental formula as potential hits (table 1). For structural annotation of unidentified peaks in GC-MS, this number of potential structures is still far too high. The idea here is to constrain these structures further by the available physicochemical information, i.e. retention time and mass fragmentation[29,52]. Gas chromatographic retention depends on boiling points as well as interaction with the column film and thus involves a variety of intermolecular forces and ultimately relies on its structural and molecular properties[53,54]. There are two alternative ways to predict retention times: either by quantitative structure-retention relationships (QSRR) methods that assess overall structural properties using chemical descriptor calculations or by additive retention contribution of individual chemical substructures without further considering interaction with other structural groups of the molecule. The NIST MS tool offers a computational prediction of retention using the latter scheme of additive contributions of chemical groups.

We have first tested the accuracy of the NIST MS tool by matching predicted versus experimental retention indices against 285 trimethylsilylated structures selected from the Fiehnlib GC-TOF MS library[42] after establishing the corresponding Kovats retention indices for these compounds. Many metabolites had grossly wrong predicted retention indices using the NIST MS group contribution tool. When inspecting errors, we found two clear dependencies on retention index errors towards trimethylsilylation. Metabolites with more than three trimethylsilylation sites showed large negative errors, i.e. predicting lower retention indices than experimentally determined (figure 3a). Terminal amines were found to be different in this regard as their retention indices showed large positive errors, i.e. underpredicting retention indices, when derivatized at both acidic protons (figure 3b). Consequently, we have used these data to implement two additional TMS group contribution rules: for terminal amines with 1x $NTMS_2$ +249 Kovats RI units were added, for 2x $NTMS_2$ +361 Kovats RI unit were used (fig. 3b). For the total number of trimethylsilyl groups derivatizing hydroxyl-, carboxyl-, thiol- and secondary amine groups, a correction formula was implemented as given by the trend line in figure 3a. The NIST MS tool is known to be also incomplete for other molecular classes (highly fluorinated compounds, cyclic siloxanes, and large or complex ring systems like steroids, adamantine analogs, and polynuclear aromatic hydrocarbons)[37] and thus, metabolites comprising these substructures can be expected to also yield incorrect retention index predictions. For non-derivatized molecules (Fig 3a and 3b), we found a range of −33 to +174 kovats RI units which is in agreement with the absolute median error value reported before[37]. After correction, a median RI prediction error of ±76 Kovats units was found for trimethylsilylated molecules, yielding a lower limit how well structure hit lists can be constrained by the improved NIST group contribution retention index prediction. Recently, Peng et al.[55] reported on the influence of steric and electronic effects on retention indices in gas chromatography. As steric descriptors can only be accurately predicted for three-dimensional optimized structures, an inclusion of this procedure was out of scope for the process detailed here.

Instead, we have applied a simple constraint for excluding elemental formulas from further consideration within the identification workflow (fig. 1): all elemental formulas were excluded for which no chemical structure was found that had a predicted retention index of at least 76 Kovats units below or above the experimental Kovats retention index. Otherwise, all structures for a given elemental formula were retained for further consideration. If mass spectrometers were used with improved mass and isotope accuracy, or if retention index prediction models were greatly improved, wrong elemental formulas and wrong structures would be excluded more precisely. Nevertheless, even this relaxed retention index filtering performed surprisingly well. In all cases except one, false elemental compositions were removed as no structure could be found in PubChem that yielded derivatized molecules with the correct accurate mass and with a range of predicted retention indices that covered the experimentally determined retention index. Even more important, only one true elemental composition was removed (pyruvate methoxime.1TMS). In effect, 36,623 structures were not considered any longer as potential hits, leaving only an average of 105 structures per elemental formula for investigation of mass fragmentations.

Prediction of molecular fragmentation requires accurate masses for constraining structures

Chemical ionization often leads to few fragment ions. The final step 6 in the metabolite annotation workflow (fig 1) therefore constrains remaining structures by matching predicted versus experimental electron ionization mass spectra with the aim to exclude as many incorrect structures as possible. Unfortunately, extensive research has shown that hard ionization leads not only to fragmentation but also to a plethora of rearrangements of radical ions[29]. We have here used MassFrontier software that generalizes and applies amassed mass spectral fragmentation pathways to chemical structures. Structures can be batch-uploaded by users in order to predict all potential ions that could be formed based on current fragmentation rules. Mass Frontier cannot predict relative ion intensities and it routinely generates many more potential fragments than experimentally found for a specific molecule.

Nevertheless, using accurate mass information, one can use and score all structures by yielding a percentage of which experimentally determined ions could not be verified by *in-silico* fragmentations in Mass Frontier. In figure 4, this approach is exemplified for three of the 66 structures tested structures that remained for $C_4H_5N_2O_3$.3TMS. While all determined accurate mass ions could be explained for the correct structure ‘asparagine.3TMS’ using a 10 ppm mass window, other potential structures yielded fragmentation pathways explaining the observed nominal masses (e.g. m/z 116, m/z 132 and m/z 141) but different accurate masses as determined for isoasparagine or 3-amino-1,4-dihydroxy-pyrrolidin-2-one. When applying this schema to all test compounds, most notably, over 89% of all remaining wrong structures could be removed by MassFrontier (table 1). In 65% of the cases that were applied to MassFrontier matching, the correct metabolite structure was retrieved within the top-3 hits (in 73% of the tested cases within the top-5 hits) and only in 2 test cases was the correct structure not found within the top-10 hits.

Conclusions

This is the first time that a comprehensive workflow was presented to annotate GC-MS peaks without library spectra matching, just using accurate masses and matching predicted against experimental retention indices and mass spectra. Given the unbiased nature of the workflow presented here, including the large structural diversity of the current 31 million unique structures deposited in PubChem, the success rate for structure annotations from database queries was remarkably high. Nevertheless, the workflow was restricted to annotate chemical structures that have been deposited before to PubChem, i.e. known to physically exist. A general process to explore all chemically possible structures for a given elemental formula would yield many more structures, and hence, retention index prediction as well as mass fragmentation predictions would need to be greatly improved for a real *de-novo* annotation scheme. Large improvements can be expected by using more rigorous retention index predictions with three-dimensional structures of trimethylsilylated metabolites and steric and electronic retention contributions of the molecule as a whole and all its atoms, instead of using a group contribution model. Adding algorithms to detect biotransformations could further aid the constraint workflow[56]. Moreover, mass spectrometers with improved mass and isotope accuracy are already commercially available that have the potential to limit the number of input molecular formulas to only the top-hit after applying the ‘Seven Golden Rules’ algorithm.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank ChemAxon for free academic licenses of Marvin, and Instant-JChem for structure handling. (https://www.chemaxon.com/forum/ftopic193.html)

Funding

Funding was provided by NIEHS grant R01 13932 to OF, support by NSF grant MCB-0820823 funded the postdoctoral scholar SK, cheminformatics work by TK was funded through NIGMS grant R24 GM78233 and the breast tumor pathology characterization by CD was provided by European Union FP7 Health-2007-2.1.4.1 project 200327.

References

1. Bowen BP, Northen TR. J Am Soc Mass Spectrom. 2010; 21:1471–1476. [PubMed: 20452782]
2. Hegeman AD. Briefings in Functional Genomics and Proteomics. 2010; 9:139–148.
3. Kind T, Fiehn O. Bioanalytical Reviews. 2010; 2:23–60. [PubMed: 21289855]
4. Fiehn O. TrAC, Trends Anal Chem. 2008:261–269.
5. Pirok G, Máté N, Varga J, Szegezdi J, Vargyas M, Dóránt S, Csizmadia F. J Chem Inf Model. 2006; 46:563–568. [PubMed: 16562984]

6. NIST/EPA/NIH. NIST Mass Spectral Library, National Institute of Standards and Technology. US Secretary of Commerce; Washington. DC, USA: 2005.
7. Wiley. Wiley Registry of Mass Spectral Data. 7. New York, USA: 2005.
8. 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]
9. Cloarec O, Campbell A, Tseng LH, Braumann U, Spraul M, Scarfe G, Weaver R, Nicholson JK. Anal Chem. 2007; 79:3304–3311. [PubMed: 17394288]
10. Akira K, Mitome H, Imachi M, Shida Y, Miyaoka H, Hashimoto T. J Pharm Biomed Anal. 2010; 51:1091–1096. [PubMed: 20007013]
11. Kolmonen M, Leinonen A, Kuuranne T, Pelander A, Ojanperä I. Drug Test Anal. 2009; 1:250–266. [PubMed: 20355204]
12. Pelander A, Ojanperä I, Laks S, Rasanen I, Vuori E. Anal Chem. 2003; 75:5710–5718. [PubMed: 14588010]
13. Halket JM, Waterman D, Przyborowska AM, Patel RKP, Fraser PD, Bramley PM. J Exp Bot. 2005; 56:219–243. [PubMed: 15618298]
14. Quirantes-Piné R, Arráez-Román D, Segura-Carretero A, Fernández-Gutiérrez A. J Sep Sci. 2010; 33:2818–2827. [PubMed: 20715141]
15. Nevedomskaya E, Ramautar R, Derks R, Westbroek I, Zondag G, Van Der Pluijm I, Deelder AM, Mayboroda OA. J Proteome Res. 2010; 9:4869–4874. [PubMed: 20690666]
16. Zhao C, Liang Y, Hu Q, Zhang T. Fenxi Huaxue. 2005; 33:715–721.
17. Neumann S, Böcker S. Anal Bioanal Chem. 2010; 398:2779–2788. [PubMed: 20936272]
18. Schymanski EL, Meringer M, Brack W. Anal Chem. 2009; 81:3608–3617. [PubMed: 19323534]
19. Fiehn O, Wohlgemuth G, Scholz M, Kind T, Lee DY, Lu Y, Moon S, Nikolau B. Plant J. 2008; 53:691–704. [PubMed: 18269577]
20. Stein SE, Scott DR. J Am Soc Mass Spectrom. 1994; 5:859–866.
21. Peng CT, Yang ZC, Maltby D. J Chromatogr. 1991; 586:113–129. [PubMed: 1806548]
22. Pierce, AE. Silylation of Organic Compounds. Pierce Chemical Co; Rockford, IL: 1982.
23. Kind T, Fiehn O. BMC Bioinf. 2007; 8:105.
24. Abate S, Ahn YG, Kind T, Cataldi TRI, Fiehn O. Rapid Commun Mass Spectrom. 2010; 24:1172–1180. [PubMed: 20301109]
25. Schymanski E, Bataineh M, Goss K, Brack W. TrAC, Trends Anal Chem. 2009; 28:550–561.
26. Garkani-Nejad Z, Karlovits M, Demuth W, Stimpfl T, Vycudilik W, Jalali-Heravi M, Varmuza K. J Chromatogr A. 2004; 1028:287–295. [PubMed: 14989482]
27. Ghavami R, Faham S. Chromatographia. 2010; 72:893–903. [PubMed: 21088689]
28. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. BMC Bioinf. 2010; 11
29. Schymanski EL, Meringer M, Brack W. Anal Chem. 2011; 83:903–912. [PubMed: 21226466]
30. Kittiratanapiboon K, Jeyashoke N, Krisnangkura K. J Chromatogr Sci. 1998; 36:361–364.
31. Héberger K. Anal Chim Acta. 1989; 223:161–174.
32. Ciazynska-Halarewicz K. Acta Chromatography. 2000; 10:56–75.
33. Farkas O, Héberger K, Zenkevich IG. Chemometrics Intellig Lab Syst. 2004; 72:173–184.
34. Kowalska T. Acta Chromatography. 2001:7–11.
35. Hemmateenejad B, Javadnia K, Elyasi M. Anal Chim Acta. 2007; 592:72–81. [PubMed: 17499073]
36. Liao L, Qing D, Li J, Lei G. J Mol Struct. 2010; 975:389–396.
37. Stein SE, Babushok VI, Brown RL, Linstrom PJ. J Chem Inf Model. 2007; 47:975–980. [PubMed: 17367127]
38. PubChem Power User Gateway (PUG). ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_pug.pdf
39. Standardizer was used for structure canonicalization and transformation. JChem 5.3.8. 2010. ChemAxon (http://www.chemaxon.com

40. Reactor was used for enumeration and reaction modeling. JChem 5.3.8. 2010. ChemAxon (http://www.chemaxon.com

41. Instant JChem was used for structure database management, search and prediction. Instant JChem 5.3.8. 2010. ChemAxon (http://www.chemaxon.com)

42. Kind T, Wohlgemuth G, Lee DY, Lu Y, Palazoglu M, Shahbaz S, Fiehn O. Anal Chem. 2009; 81:10038–10048. [PubMed: 19928838]

43. Eisenberg F Jr, Bolden AH. Anal Biochem. 1969; 29:284–292. [PubMed: 5792565]

44. Fiehn O. TrAC, Trends Anal Chem. 2008; 27:261–269.

45. Katona ZF, Sass P, Molnár-Perl I. J Chromatogr A. 1999; 847:91–102.

46. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T. J Mass Spectrom. 2010; 45:703–714. [PubMed: 20623627]

47. Song Y, Aglipay JA, Bernstein JD, Goswami S, Stanley P. Cancer Res. 2010; 70:3361–3371. [PubMed: 20395209]

48. Harrison, AG. Chemical Ionization Mass Spectrometry. 2. CRC Press; Boca Raton: 1992.

49. Kind T, Fiehn O. BMC Bioinf. 2006; 7

50. Monroe, M. MWTWIN V635. 2005. [http://alchemistmatt.com/]

51. Fiehn O, Kopka J, Trethewey RN, Willmitzer L. Anal Chem. 2000; 72:3573–3580. [PubMed: 10952545]

52. Hill DW, Kertesz TM, Fontaine D, Friedman R, Grant DF. Anal Chem. 2008; 80:5574–5582. [PubMed: 18547062]

53. Hanai T, Hubert J. J Chromatogr A. 1984; 302:89–94.

54. Zhang X, Lu P. J Chromatogr A. 1996; 731:187–199.

55. Peng CT. J Chromatogr A. 2010; 1217:3683–3694. [PubMed: 20227699]

56. Liotta E, Gottardo R, Bertaso A, Polettini A. J Mass Spectrom. 2010; 45:261–271. [PubMed: 20014151]

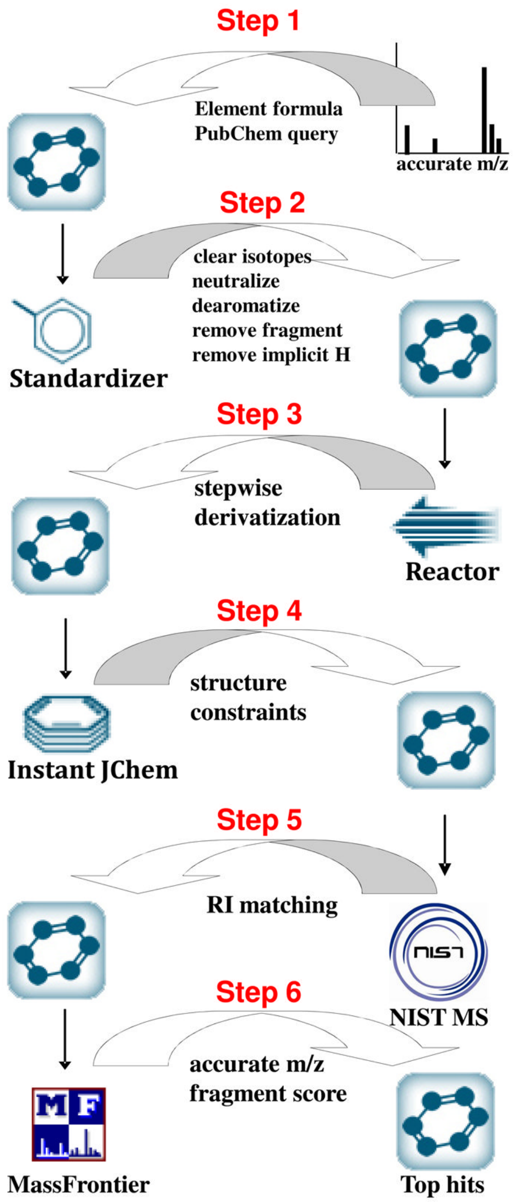

Figure 1.
Workflow for *de-novo* annotation of unknown metabolic peaks matching predicted *versus* experimental data in derivatization based accurate mass GC-TOFMS analysis.

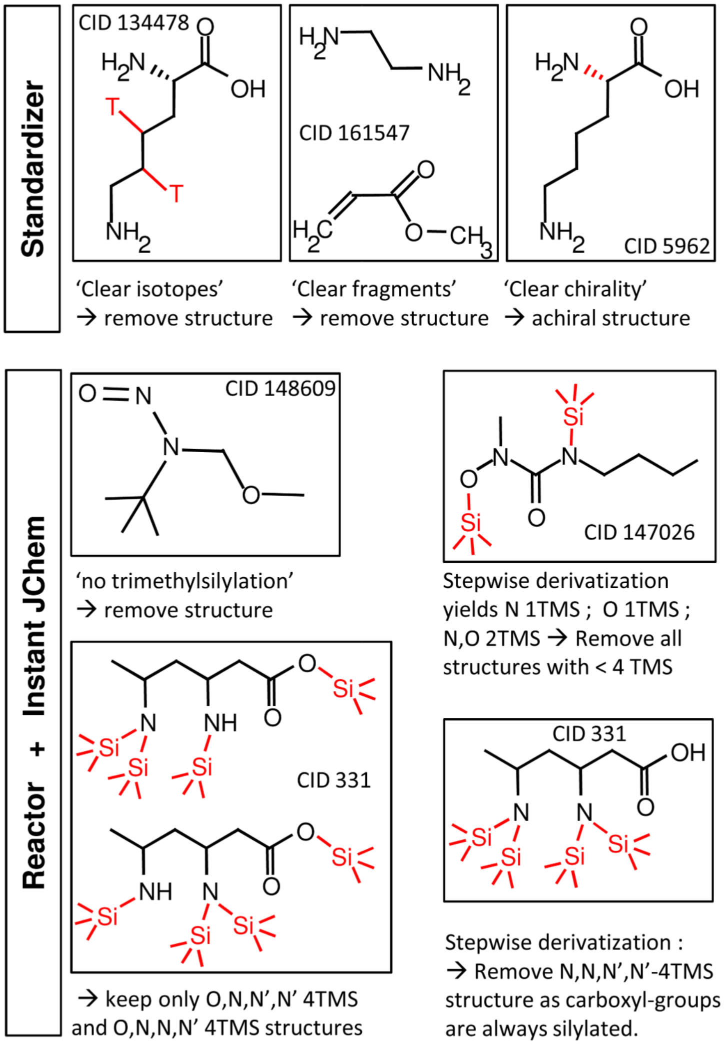
Standardizer
CID 134478
'Clear isotopes'
→ remove structure
CID 161547
'Clear fragments'
→ remove structure
CID 5962
'Clear chirality'
→ achiral structure
Reactor + Instant JChem
CID 148609
'no trimethylsilylation'
→ remove structure
CID 147026
Stepwise derivatization yields N 1TMS ; O 1TMS ; N,O 2TMS → Remove all structures with < 4 TMS
CID 331
→ keep only O,N,N',N' 4TMS and O,N,N,N' 4TMS structures
CID 331
Stepwise derivatization :
→ Remove N,N,N',N'-4TMS structure as carboxyl-groups are always silylated.

Figure 2.

Constraining structures using the JChem 'Standardizer' and 'Reactor' tools. Example for 6 out of 570 structures downloaded from the PubChem database for query of C6H14N2O2, constrained as tetra-trimethylsilylated structure. Standardizer: Structures CID 134478, 161547 are removed due to isotope inclusion or fragmented structures. Chiral centers are removed. Reactor: Structures CID 148609, 147026 are removed as all derivatization products have less than 4 trimethylsilyl groups (TMS). Structure CID 331 yields 2 new structures, each with 4 TMS groups which are used for retention index calculation.

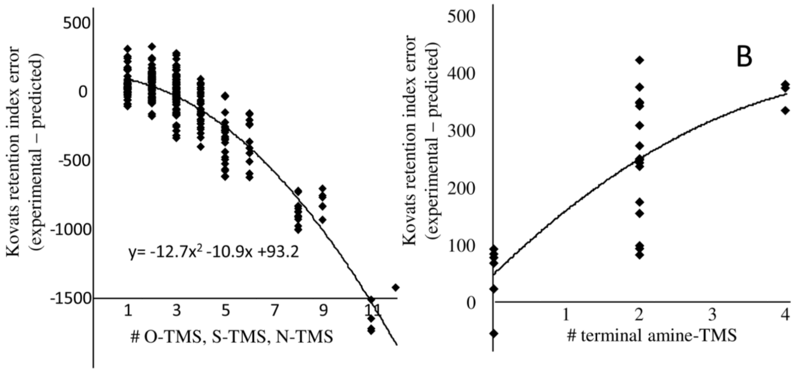

Figure 3. TMS-group contribution error for retention index prediction using the NIST MS tool

Figure 3A: Difference of experimental versus predicted spectra for 286 hydroxyl-, carboxyl-, amine- and thiol-comprising metabolites of the Fiehnlib metabolomics library. The trendline for dependence of prediction error versus number of trimethylsilyl groups was used as additional TMS-group contribution correction.

Figure 3B: 26 trimethylsilylated terminal amines from the Fiehnlib metabolomics library showed a positive trend from underivatized amines to tetraTMS-derivatized amines. This error for 2 and 4 terminal amine-TMS groups was subsequently used to correct the NIST MS tool as additional group contribution.

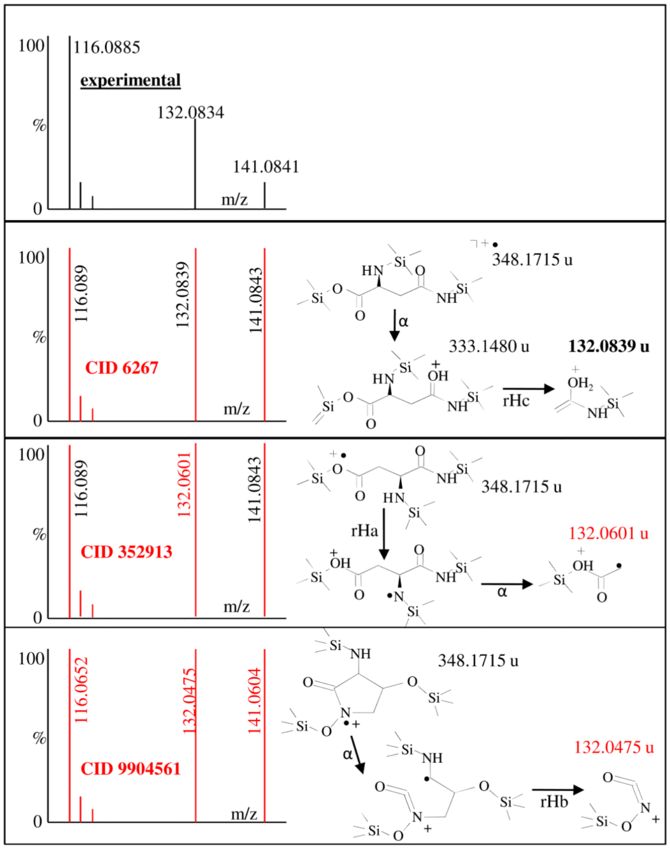

Figure 4.

Scoring remaining structures by accurate mass EI-spectra matching in Mass Frontier. Top panel: experimental spectra (here magnified from m/z 116–141 u) are matched against all ions that are predicted by the Mass Frontier algorithm (lower panels, spectra labeled red). For clarity, computational fragment ions that are not matching at least the nominal experimental masses are left out. For asparagine 3TMS (PubChem CID 6267), all experimental ions were computationally predicted within 10 ppm mass range. For CID 352913 (isoasparagine 3TMS), and CID 9904561 (3-amino-1,4-dihydroxy-pyrrolidin-2-one 3TMS), no ions matching m/z 132.0834 could be found (labeled in red) when considering all possible fragmentation routes given by Mass Frontier.

Table I

Top-3 hits for elemental composition of 5 examples of 29 test metabolite peaks determined by ammonia chemical ionization - accurate mass GC-TOF MS and constraining all potential structures by retention index prediction. True positives are hits are marked in red, false positives and false negatives are marked in blue. The full data set is given as supplement 4.

Test metabolite derivatization status mass error, isotope A+1/A; A+2/A error	top-3 sum formulas with PubChem query	# non-derivatized PubChem structures	#structures curated + derivatized	TMS-corrected NIST prediction of Kovats RI		experim. Kovats RI	RI constraint	Mass-Frontier rank #
				minimal	maximal			
nicotinate 1TMS	$C_9H_{13}NO_2Si$	113	32	1038	1551	1257	TRUE	2
−13.2ppm	$C_{10}H_{13}NOS$	517	163	1465	1910		FALSE	
+2.7%; +0.5%	$C_{13}H_9NO$	223	132	1415	1984		FALSE	
tocopherol 1TMS	$C_{32}H_{58}O_2Si$	265	28	2698	3460	3112	TRUE	3
+5.9 ppm	$C_{36}H_{54}O$	20	15	3243	3964		FALSE	
+5.0%; −0.4%	$C_{35}H_{54}N_2$	22	17	3515	4092		FALSE	
aspartate 3TMS	$C_{17}H_{27}NO_3Si_2$	1729	340	1730	2296	1489	FALSE	
+ 4.0 ppm	$C_{13}H_{31}NO_4Si_3$	144	21	1233	1644		TRUE	7
−1.4%; −1.0%	$C_{18}H_{27}NO_2SiS$	799	287	2074	2773		FALSE	
valine 2TMS	$C_{12}H_{27}NOSiS$	110	56	1426	1984	1185	FALSE	
−6.3 ppm	$C_{11}H_{27}NO_2Si2$	579	116	1011	1508		TRUE	3
−2.0%; −0.8%	$C_{15}H_{23}NOSi$	1885	747	1141	2040		TRUE	
pyruvate 1TMSmeox	$C_7H_{15}NO_3Si$	53	2	1057	1232	1035	FALSE	2
+3.4 ppm	$C_8H_{15}NO_2S$	277	28	1391	1789		FALSE	
+1.3%; +0.2%	$C_5H_{11}N_5O_3$	35	0	*nd*	*nd*		FALSE	