Published in final edited form as:
Anal Chem. 2009 December 15; 81(24): 10038–10048. doi:10.1021/ac9019522.

FiehnLib – mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry

Tobias Kind, **Gert Wohlgemuth**, **Do Yup Lee**, **Yun Lu**, **Mine Palazoglu**, **Sevini Shahbaz**, and **Oliver Fiehn**
University of California Davis, Genome Center, Davis, CA 95616, USA

Abstract

At least two independent parameters are necessary for compound identification in metabolomics. We have compiled 2,212 electron impact mass spectra and retention indices for quadrupole and time-of-flight GC/MS for over 1,000 primary metabolites below 550 Da, covering lipids, amino acids, fatty acids, amines, alcohols, sugars, amino-sugars, sugar alcohols, sugar acids, organic phosphates, hydroxyl acids, aromatics, purines and sterols as methoximated and trimethylsilylated mass spectra under electron impact ionization. Compounds were selected from different metabolic pathway databases. The structural diversity of the libraries was found to be highly overlapping with metabolites represented in the BioMeta/KEGG pathway database using chemical fingerprints and calculations using Instant-JChem. In total, the FiehnLib libraries comprised 68% more compounds and twice as many spectra with higher spectral diversity than the public Golm Metabolite Database. A range of unique compounds are present in the FiehnLib libraries that are not comprised in the 4,345 trimethylsilylated spectra of the commercial NIST05 mass spectral database. The libraries can be used in conjunction with GC/MS software but also support compound identification in the public BinBase metabolomic database that currently comprises 5,598 unique mass spectra generated from 19,032 samples covering 279 studies of 47 species (plants, animals and microorganisms).

Introduction

Compound identification in metabolomics

Metabolomics is the unbiased detection and quantification of small molecules that are transformed by (bio)chemical reactions in cells.[1] Species certainly exert different numbers of metabolites, e.g. comprising secondary natural products that may serve specific roles in the physiology, communication or defense of organisms. In addition, many metabolites are shared across organisms, serving general functions in energy metabolism, protein and DNA biosynthesis or cell wall and membrane stability. Therefore, the structural diversity of metabolites is extraordinarily large. Due to this chemical complexity, no single technology platform is capable to profile all metabolites simultaneously. Instead, metabolomics technologies strive to detect, quantify and identify as many compounds with as few methods and platforms as possible. For this reason, platforms need to be universal in detection with high sensitivity and excellent compound specificity. Mass spectrometry is an excellent tool fulfilling these requirements, mostly used in combination with liquid or gas phase chromatography to

Correspondence to: Oliver Fiehn.

Competing interests
The FiehnLib libraries are commercially available under Agilent #G1676AA and Leco #359-008-100 under license agreements with UC Davis.

enhance the compound specificity and the number of compounds detected. Gas chromatography-coupled mass spectrometry (GC/MS) has been used for metabolite profiling decades before liquid chromatography/mass spectrometry (LC/MS),[2,3] primarily due to the difficult ionization process in LC/MS that demands compounds to be ionized, separated from the liquid solvents and subsequently funneled into the mass spectrometer for further analysis. Mostly, electrospray ionization is used in LC/MS which is inherently softer than GC/MS electron impact ionization, generating mostly molecular ion adducts with no or little in-source fragmentation. Conversely, electron impact ionization is a hard ionization that has been standardized at 70 eV in the late 1960's, causing ionization of neutral molecules with subsequent immediate fragmentation and rearrangement reactions that yield complex mass spectra that are specific for each molecule.

Consequently, libraries of mass spectra[4] as well as software programs[5] to search those spectra have been compiled over the past decades[6] to aid compound identification through mass spectral matching. Mass spectra of stereoisomers and positional isomers often yield nearly identical spectra and are therefore not unique enough for unambiguous compound identification. Furthermore, the hard electron impact ionization at 70 eV often results in highly abundant generic fragment ions but low abundant or missing ions for larger fragments or even molecular ions. Therefore, the information of standardized chromatographic retention times, called retention index, must aid the identification of compounds. Retention indices have been introduced by Kovats based on aliphatic carbon numbers.[7] Nevertheless, only recently a comprehensive mass spectral/retention index database has been compiled by NIST.[8] In metabolomics, alkanes are less useful for retention index generation because these may be part of the metabolome to be analyzed, e.g. in biofuel applications. Alternatively, retention indices might be calculated from molecular properties and a group contribution algorithm for gas chromatographic retention index calculation was created.[9] Due to the number of different column types, column lengths, phase ratios and phase selectivities, the algorithm itself is not accurate enough for single isomer identification but can be used as a powerful retention index filter.

While the quality and comprehensiveness of retention index/mass spectral libraries is critical, it is equally important to obtain pure spectra from metabolomic samples. As GC/MS chromatograms can be very complex, chromatographic peaks will be heavily overlapping. Pure spectra can be obtained by deconvoluting overlapping spectra and noise subtraction. A recent review discussed the impact of parameter selection on the deconvolution packages AMDIS, ChromaTOF and AnalyzerPro to minimize mass spectral deconvolution errors, an important quality aspect when using the FiehnLib libraries.[10] Among several mass spectral and retention index collections specialized for GC-MS based metabolic profiling, the publicly available Golm Metabolomic Database (GMD) has been frequently used.[11] For LC-MS based metabolomics, MassBank12 and Metlin13 were released as online services. Other databases containing both mass spectra and retention indices are the RIZA GC-MS database14, the Adams collection of volatile compounds[15] or the MASSFinder library of natural compounds and terpenoids.[16]

In this article we present the FiehnLib GC/MS libraries that are based on a fatty-acid methyl ester retention index system that is in use since 2003. The libraries have been established by GC/MS based on time-of-flight mass spectrometry (GC-TOF) and based on quadrupole mass spectrometry (GC-Quad) using similar, but not identical, chromatographic and mass spectrometric parameters. Structural overlap analysis with available metabolomic and biochemical databases demonstrates the comprehensiveness of the libraries. The use of FiehnLib is shown for individual chromatograms using mass spectrometry software, as well as for BinBase a public GC-TOF metabolomic repository.[17]

Methods

Data acquisition

All metabolite reference standards underwent a two-step derivatization procedure. Therefore 1 mg of each standard was dissolved in a solution of 1 ml methanol:water:isopropanol (2.5:1:1 v/v). Then 10 µl of each standard solution were taken out and evaporated to dryness. First, methoximation was performed to inhibit the ring formation of reducing sugars, protecting also all other aldehydes and ketones. A solution of 40 mg/ml O-methylhydroxylamine hydrochloride, (CAS: [593-56-6]; Formula CH5NO.HCl; Sigma-Aldrich No. 226904 (98%)) in pyridine (99.99%) was prepared. The dried standards and 10 µl of the O-methylhydroxylamine reagent solution were mixed for 30 s in a vortex mixer and subsequently shaken for 90 minutes at 30°C. Afterwards, 90µl of N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) (1 ml bottles, Pierce, Rockford IL) was added and shaken at 37°C for 30 min for trimethylsilylation of acidic protons to increase volatility of metabolites. A mixture of internal retention index (RI) markers was prepared using fatty acid methyl esters (FAME markers) of C8, C9, C10, C12, C14, C16, C18, C20, C22, C24, C26, C28 and C30 linear chain length, dissolved in chloroform at a concentration of 0.8 mg/ml (C8-C16) and 0.4 mg/ml (C18-C30). 2 µl of this RI mixture were added to the reagent solutions, transferred to 2 mL glass crimp amber autosampler vials. Data acquisition parameters are given in table 1. Subsequent to data processing using the instrument manufacturer's software programs, spectra and retention indices were manually curated into the new Leco FiehnLib (359-008-100) or automatically transferred by Agilent to the new Agilent FiehnLib (G1676AA).

Chemical structure overlaps between BioMeta/KEGG, LipidMaps, FiehnLib and GDB

BioMeta SD files18 were downloaded from the BioMeta website19 in version of June 8th, 2007. The LipidMaps structure database SD files were downloaded from the LipidMaps website20 in version of Jan 17, 2007. For structure searches and overlap calculations, the free Instant-JChem database[21] was used to import and curate each single molecular structure and to assign names and check PubChem compound identifiers (CID). Additional curation was performed within Microsoft EXCEL.

The SMARTS format ('smiles arbitrary target specification') was used to match substructures of chemical structure files after conversion between SDF and SMILES file formats ('simplified molecular input line entry specification' in ASCII). SMARTS patterns were obtained from the freely available OpenBabel project.[22] We have written a Java application interface to match these SMARTS functional groups to the SD files in our Instant-JChem database. Direct counts of substructure presence in FiehnLib and BioMeta/KEGG files were used to generate table 2. For multivariate principal components analysis (PCA), chemical fingerprints were generated into a new standard Derby database table using ChemAxon's Chemical Hashed Fingerprint Generator *GenerFP* (v3.2.11), with standard fingerprint settings (16×4 bytes or 512 bits, 2 bits pattern length, 6 bonds) duplicate filtering was set. This Derby database was used for multivariate PCA statistics in unsupervised mode for dimension reduction purposes using the Statistica DataMiner vs.7.0 (StatSoft, Tulsa, OK) on unit variance scaled data.

Two Golm GC/MS database sets were downloaded in August 2007: annotated GC-TOF-MS spectra23 (T_MSRI_ID; 2004-03-01, curated) and annotated quadrupole GC-MS spectra (Q_MSRI_IDl; 2004-03-01, curated)24. The T-MSRI DB contains 326 annotated or identified mass spectra including 1-TMS, 2-TMS or 3-TMS spectra of the same compound which refer to 306 unique compounds. Additionally the database contained 60 mass spectra from 41 unique compounds that were annotated as "match". The Q-MSRI DB contained 399 unique mass spectra from 327 unique compounds and additionally 257 mass spectra from 151 unique

structures that were annotated as matched compounds. From a regular expression analysis using all names from identified, annotated and matched spectra it can be concluded that around 603 unique compounds were under investigation. The Visual C++ DLL code was obtained from NIST and a search algorithm was implemented from existing templates using the freely available Microsoft Visual C++ compiler. Spectra were compared starting from 80 m/z using the dot product algorithm.25 Additionally the NIST mass spectral search algorithm26 implemented in JAVA was downloaded to perform similarity searches based on cosine correlation of peaks. The algorithm provides comparable search results to the NIST direct and reverse search which result in a hit statistics from a threshold level (set to 0.5; low similarity) to 1.0 (direct hit; best similarity). The correlation threshold was set to 0.5, the correlation type to cosine, a squared weighing factor was selected, normalization was performed with square root and a cutoff threshold of 20 was used.

Computer Hardware

All calculations were performed under Windows XP on a MonarchComputer Dual-Opteron 254 (2.8 GHz, 2.8 GByte RAM), equipped with an Areca ARC-1120 RAID-6 array using Western Digital Raptor SATA hard disks. A RAID-6 array provides fault tolerance from two drive failures. This equipment enabled hard disk burst read-write transfer rates of more than 500 MByte/s. An additional RamDisk (QSoft Ramdisk Enterprise) was used for file based operations allowing read-write rates of up to 1000 MByte/s.

Results

Selection of relevant substances from the KEGG biochemical pathway database

Metabolite reference standards were selected according to their presence in the KEGG biochemical LIGAND database[27] using a mass cut-off at 600 Dalton. High molecular mass molecules are not readily volatilized, even as trimethylsilylated derivatives and therefore they are unsuitable for GC/MS analysis. Matching KEGG structure files against website query tools for vendors of chemical substances proved difficult as vendors regularly list their compounds as salt adducts, tautomers or without adequate synonym lists and cross-references. Matching compounds by name tags was difficult due to the high number of synonyms available, and similarly difficult was an accurate and comprehensive search by chemical identifiers (CAS or PubChem ID). When the FiehnLib compounds were purchased in 2005, SD file downloads and web engines were uncommon. A recent query using ChemNavigator's iResearch system using a near match search (which includes salt and stereochemistry variants) revealed a commercial availability of 4,117 hits of around 10,000 compounds from the cleaned KEGG/BioMeta database. For FiehnLib, KEGG structures were matched against the MDL Available Chemicals Directories Oracle concordance database. A list of MFCD chemical catalog numbers was obtained as a result, to compare chemical prices from different vendors. Metabolites above $200 per reference standard were usually excluded, yielding around 1,200 chemicals purchased mostly from VWR, Fisher Scientific, Sigma-Aldrich-Fluka and Steraloids Inc. Each reference compound was individually derivatized and analyzed by GC/MS as given in the methods section. Ketone and aldehyde groups are first protected by reaction with O-methylhydroxylamine hydrochloride to form two methoxime isomers (meox, Z/E or syn/anti, figure 1a) which is especially important for carbohydrate analysis28 to inhibit formation of multiple peaks from open-chain and cyclic sugar anomers.[29] In a second reaction step, acidic protons are exchanged against trimethylsilyl derivatives (TMS) by MSTFA to increase volatility. Each chromatogram was manually curated to match acquired spectra against expected fragments and molecular ions of methoximated and trimethylsilylated derivatives. For the example of estrone (figure 1), the molecular mass of the underivatized molecule (M=270 Da) was increased by addition of one methoxime group (+29 Da) and one trimethyl group (+72 Da) yielding an abundant molecular ion of 371 Da with the typical fragmentation

of M-15 for methyl cleavages and a complex pattern of fragmentation and rearrangement ions. While the two methoxime derivatives yield near-identical mass spectra, differences in retention due to different steric parameters are usually sufficient to yield two chromatographically resolved GC/MS peaks.

Conversion of retention times to Fiehn-retention index (RI) and to Kovats-RI

Absolute retention times differ based on column conditions. Despite using the Agilent retention-time locking system,[30] we found differences in absolute retention times of up to 24 s for low-boiling compounds when using split versus splitless injections. As gas chromatographic conditions were not isothermal, it is advisable to use a large range of internal marker compounds to obtain a retention index with fix numbers that is static over time. The classic Kovats-RI uses alkanes as internal retention markers. Currently 293,247 Kovats RI values for 44,008 compounds are collected in the new NIST08 retention index database, also representing the frequent use of the Kovats index in the peer-reviewed literature. However, apart from the presence of alkanes in plant and biofuel samples, we have found the spectral properties of fatty acid methyl esters (FAMEs) to be superior for unambiguous and automated detection of the retention index markers in complex samples. In order to compare different mass spectral/retention index libraries, RI numbers[31] need to be convertible.32 When analyzing the Fiehn FAME retention index standards concomitant with the Kovats series of alkanes, we found a 5^{th} grade polynomial equation sufficient interconvert the retention indices with an adjusted R^2=0.99944 and a standard error of estimate s=20.77 Kovats units. The obtained formula was specific for the GC conditions given in the methods section and can be found in the supplement section.

Curation of molecular structures and mass spectra from the FiehnLab mass spectral database

Structures of FiehnLib metabolites were obtained in structure data file format (SDF) from NCBI's PubChem chemical database via hyperlinks using a JAVA interface for the PubChem Power User Gateway (PUG). PUG is an XML-based interface to PubChem that can perform structure search and structure download and includes a SOAP-compliant wrapper.[33] The initial 1,162 GC-TOF mass spectra were manually curated and a methoximated and trimethylsilylated structure was assigned using the ChemAxon Reactor program.34 For 143 structures no correct structure of the TMS compound could be assigned due to complex rearrangement reactions or missing molecular ions. 797 structures were unambiguously assigned. For the remaining 222 structures, TMS/methoximated derivatives were assigned with high probability based on chemical reactivities.

Structures of biochemical pathway intermediates were obtained from the BioMeta database35 which is a structurally curated version of the LIGAND database in KEGG.27 Using the external 'Standardizer' graphical user interface in Instant-JChem,[21] structures were aromatized, explicit (lonely) hydrogens were removed as well as structure fragments (keeping the largest), structures were tautomerized, mesomerized and absolute stereo configurations were removed. All Markush structures from the BioMeta DB were removed using molconvert and converted to SMILES format using aromatization without explicit hydrogens. Additionally all wildcard atoms containing "*" were removed from the SMILES file. Stereochemistry search was turned off. After standardizing, 11,280 BioMeta-curated KEGG metabolites structures and 701 unique compounds of the Leco FiehnLib were used for overlap analysis. A similar number of compounds and spectra were established for the Agilent GC-Quad FiehnLib. Therefore, chemical overlap calculations were only performed for the GC-TOF spectra.

Calculation of the chemical overlap between KEGG and LipidMaps compared to FiehnLib

The internal overlap after querying all structures of the FiehnLib in the BioMeta/KEGG database resulted in 564 compounds which were found in the BioMeta database. The LipidMaps database itself has 750 assigned KEGG IDs. Some of these non matching structures were sugar phosphates, lipids and steroids. Lipid compounds play an important role in organisms; therefore the LipidMaps consortium assembled a large structure database containing 7,946 unique structures. Only 861 lipid compounds (11%) from LipidMaps could be found in the KEGG/BioMeta DB. The FiehnLib had an overlap of 80 compounds (11%) with LipidMaps. Based on a molecular weight cut-off of 500 Dalton, the BioMeta DB contains 9,234 compounds and LipidMaps contains 2,887 compounds. The FiehnLib covers 19% of small lipid-like molecules and 6% of the BioMeta substance space. The final numbers might change with recent database updates.

Apart from direct hits, it was important to investigate potential bias and structural complexity of FiehnLib compounds in comparison to KEGG metabolites. The workflow for chemical structure comparisons is shown in figure 2. To visualize the overlap of structural diversity between the BioMeta database (KEGG) and the substances in the FiehnLib database, we generated a bit fingerprint for each compound.[36] Compound bit fingerprints comprise structural information in “0” and “1” bits for use in chemical database handling and combinatorial chemistry. As our libraries are small compared to the combinatorial chemical libraries used in pharmaceutical research, 16 integer fingerprints were found sufficient for principal component analysis (PCA). The fingerprint data matrices were used as variables in unsupervised PCA dimension reduction for visual comparison of structural complexity and library overlaps. The graphics of the principal component scores plots (figure 3) demonstrate that the 11,280 curated KEGG metabolite structures were highly diverse. Overlapping dots represent the same component. In comparison, the FiehnLib structures showed a very similar density of structures, with no clear clustering and an equally diverse range of structures, covering even distant and sparsely populated parts of the graph. This finding confirms that compounds comprised in FiehnLib are equally diverse as biochemicals that constitute the KEGG library.

For a more detailed investigation of relative coverage of discriminant functional groups and substructures, we performed a diversity calculation using freely available SMARTS patterns using 303 common functional groups from the OpenBabel project.[22] A matrix containing the hit count for each SMARTS match was generated (table 2). The quantitative comparison of functional groups that are comprised in the BioMeta/KEGG DB and the FiehnLib compounds shows that both major and minor representatives of structural moieties are equally represented. This finding proves that no single compound class is overrepresented. The only missing structures in the FiehnLib are chlorines and alkynes which were intentionally disregarded when chemical reference standards were purchased due to limited relevance in biochemical pathways.

Overlap calculation between the NIST05, GolmDB and FiehnLib mass spectral databases

While the NIST05 and the FiehnLib mass spectral libraries are assigned by compound structures, no structural information or unique PubChem CIDs were downloadable for the public GolmDB quadrupole and TOF libraries. Using a regular expression analysis of the compound names comprising the GolmDB libraries, it was concluded that the public TOF GolmDB comprises approximately 347 unique compounds (386 spectra) and the public Quadrupole GolmDB consists of around 478 compounds (656 spectra) with a combined total of around 603 compounds. Using unique PubChem compound identifiers, the FiehnLib libraries comprise 714 metabolites (1,050 spectra) for the Agilent GC-Quad library, 722 metabolites (1,162 spectra) for the Leco GC-TOF library and a combined total of 1,014 compounds for which structures, PubChem CIDs and further information are freely

downloadable.[37] These numbers indicate that the FiehnLib libraries cover around 1.5–2 fold more metabolites and twice as many spectra as the GolmDB GC/MS libraries.

Apart from comparing the number of unique compounds and spectra, we investigated the spectral diversity. For overlap visualization, mass spectra were transformed into compressed mass spectral features using Varmuza's MassFeatGen program using prominent peak abundances as well as features like modulo-14 computations.38 Mass spectral features are more useful for chemometric analysis than raw mass spectra because the features inherently comprise structural information such as combinations of fragmentation patterns and peak abundances[39] and zero values are removed. Unsupervised principal component analysis (PCA) of all mass spectral features demonstrates that the GC-TOF FiehnLib has more mass spectra and that its spectral diversity space is larger than the GolmDB (figure 4). However a large number of spectra are overlapping because both libraries contain important sugars and amino acids. The NIST05 database was purchased and the freely available NIST MS Search program[40] was used to extract only compounds with TMS structures. 4,345 spectra were retrieved and used for spectral comparisons. The spectral space of NIST05 trimethylsilylated compounds was even larger than the FiehnLib mass spectra (see supplemental data S2), which is due to the high number of xenobiotic compounds in the NIST05 library, from pesticides to drugs and industrial chemicals. Nevertheless, many unique compound spectra were found in the FiehnLib mass spectral library due to the fact that multiple compounds are both methoximated and silylated, whereas in drug and forensics analysis compounds are usually only trimethylsilylated.

Use of FiehnLib in routine metabolic profiling for BinBase database queries

The FiehnLib libraries can be used in conjunction with instrument vendor software such as Agilent's ChemStation or Leco's ChromaTOF software. When applying the FiehnLib Agilent GC-Quad library, 102 compounds were unambiguously identified by mass spectral and retention time matching in human blood plasma after mass spectral deconvolution with AMDIS. For GC-TOF data, manual investigation of chromatograms may retrieve identified compounds that are missed in automatic screens. All GC-TOF profiles generated at the UC Davis Genome Center's metabolomic core and research laboratories are processed, stored and disseminated via the open source SetupX/BinBase metabolomics database system41 which handles sample organization, class assignments and provides downloadable web services for results. Importantly, the BinBase database [17] performs automatic processing of both known and novel metabolites independent of the sample origin, as it is based on the identical FAME retention index markers and data acquisition parameters as the FiehnLib GC-TOF library. Currently, BinBase includes 5,598 unique mass spectra generated from 19,032 samples covering 279 studies of 47 species (plants, animals and microorganisms). Over all GC-TOF samples processed by BinBase, 560 deconvoluted spectra were detected per sample. On average 47% of these spectra were removed by the consistency filtering algorithm implemented in BinBase, leaving 310 metabolites in result files of which an average of 116 compounds were identified. Identifying spectra in BinBase is supported by all FiehnLib libraries as well as the NIST05-trimethylsilylated sub-library. So far, 677 compounds in BinBase have been identified by names linked to PubChem CIDs of which only 90 were annotated on high mass spectral similarity to NIST05 entries. The metabolomics standard initiative (MSI) published recommendations on minimum reporting standards,[42] suggesting such NIST05 metabolites cannot be called 'identified' but only 'annotated' and need to be verified by purchasing further reference chemicals and matching retention indices in addition to mass spectra. After publication in peer-reviewed journals, BinBase data including mass spectra are regularly published for open downloads.41 In addition, user mass spectra in different formats, individual compounds or unknown spectra with BinBase identifiers can be pasted and queried through a public web GUI.37 Because the FiehnLib GC/MS libraries and BinBase are based on retention

indices, peaks can be tracked and identified in sets of very different organs, unlike less useful approaches that are based on chromatogram alignments.

To provide an example for the use of the FiehnLib libraries in conjunction with SetupX/BinBase data processing, we have queried 846 samples covering different studies in mammalian tissues and biofluids, vascular plants and microorganisms. In these combined studies, 324 unique metabolites were automatically identified by retention index/mass spectral matching. Metabolite profiles were compared qualitatively and quantitatively by calculating mean intensities for all metabolites and each organism or biofluid, and subsequently contrasted by principal component analysis (figure 5). The PCA graph demonstrates that fundamental differences between taxonomic kingdoms (animal versus plants and microorganisms) are even larger than the known abundant metabolic variations within those kingdoms (mammalian tissues and body fluids). The largest difference (vector 1, explaining 23.8% of the total variance) was spanned by the extreme metabolomes of mouse urine and algae metabolism (*Chlamydomonas reinhardtii*). Photosynthetic tissues of vascular plants spanned another extreme phenotype (vector 2 and 3). However, a closer investigation showed that even mammalian phenotypes were quite distinct in both identity and magnitude of observed metabolites as given in a Venn-Diagram (figure 5). 251 out of the total 324 positively identified metabolites were present in the mammalian body fluids blood, urine and intestinal fluids. However, only 90 metabolites were common between these fluids, with other metabolites identified as being specific to one or two organs. The identifications validated the specificity of BinBase processing, as, for example, cholic acid is a bile acid known to be specific for the intestinal physiology. Glycine adducts such as hexanoyl-glycine are typical excreted urine metabolites, and dicarboxylic acids are known to be catabolites after oxidative stress in lipids, transported by the blood stream. We propose that such comparisons can yield interesting facts about presence and absence of metabolites for pathway reconstructions, and can also be used for analyzing relatedness of species, e.g. in phylogenetic trees.

Discussion

Data acquisition for FiehnLib libraries

The selection, structural curation, pricing and ordering of metabolic reference libraries was not straightforward and still presents a challenge to generate a much larger but still biochemically relevant metabolome mass spectral library. When the FiehnLib metabolites were purchased, the only available online product websites were MDL's Available Chemicals Directory, ACS SciFinder and CambridgeSoft's ChemFinder.[43] Today, the ZINC database[44] offers chemicals for pharmaceutical research including a Natural Products Database. The eMolecules.com library45 is a database for commercial and bulk chemicals, ChemSpider46 links its entries to major chemical vendors and ChemNavigator's iResearch library47 covers more than 50 million commercially available chemicals. Nevertheless, assigning and constraining these compound lists is still a formidable task for which significant resources need to be allotted. Summing up chemical purchases, data acquisition and staff time for curation of the FiehnLib libraries, total costs of about $250,000 or around $200 per compound can be estimated. This is in agreement with Stephen R. Heller's cost calculations for mass spectra.[48] Commercial libraries such as the highly curated NIST05 mass spectral library (190,825 spectra) therefore perfectly complement the FiehnLib libraries. However, the NIST05 spectra were acquired on a large variety of instruments and chromatographic conditions and can therefore not be used for unambiguous assignment of retention indices and are not compliant to Metabolomics Standards for compound identification.[49]

Derivatization assignments of FiehnLib metabolites

Finding and assigning metabolic derivatives in GC/MS chromatograms required substantial chemical and mass spectroscopic experience that could not be carried out by undergraduate students but only by trained personnel. For example, depending on the steric hindrance of the final molecule and the thermodynamic and kinetic control of the reaction, certain chemical functional groups are sometimes not or only partially derivatized. This results in multiple peaks per compound which are separated on the gas chromatographic column. For the GC-TOF FiehnLib library, methoxime derivatives were arbitrarily called 'major' and 'minor', as exemplified in figure 1 for the compound estrone as it is not straightforward to assign the correct methoxime-E,Z position by quantum chemistry.[50,51] For a very few cases of multiple methoxime groups, up to eight chromatographic peaks were observed. Similarly, the correct status of trimethylsilylation is difficult to assess when molecular ions are absent under electron impact ionization. Primary amino groups may be derivatized once, twice or are left underivatized, depending on the acidity of the amino protons and, to our experience, on the cleanliness of the syringe, liner, injector body, column and sample matrix contributions.[52] We could not confirm reports that detailed different trimethylsilylation ratios depending on reaction times,[53] e.g. while waiting for injection on autosampler racks. The actual ratio of synthesis and degradation of trimethylsilyl derivatives may rather depend on the presence and activity of catalytic sites in the GC/MS injector and can be controlled by careful maintenance procedures. For the GC-Quad FiehnLib library, we have named derivatives by increasing numbers according to retention index, e.g. serine 1, serine 2 and serine 3 (for the derivatives with no, one or two TMS-groups derivatizing the primary amino group.)

Substructure classifications for library overlap analyses

Substructure classifications are still under active research in various research areas. Markush structures support queries in pharmaceutical research. Structures in the online CRC Dictionary of Natural Products[54] lists around 300 generic substance classes and could support substructure scaffolds using maximum common substructure approaches (MCS) or hierarchical scaffold classification.[55] New software programs[56] might be used to find discriminative fragments.[57] Further classification schemas are provided by the Medical Subject Headings Catalog (MeSH)[58] and LipidMaps[20] which suggested nomenclature for 81 general lipid classes and 276 sub-categories. Recently the KEGG BRITE database[59] was introduced which represents a collection of hierarchical classifications for compounds, enzymes and metabolic pathways. It is important to provide publicly available unique identifiers for databases and libraries that can be used to cross-reference compounds to scientific reports.[60] We suggest the use of PubChem identifiers as these can be batch downloaded and the use of InChI structure hash keys (InChIKey) for comparisons. The Chemical Abstract Service Registry Number (CAS-RN) is a good source for chemical literature references in the CAS database itself and commercial chemicals but not for batch compound queries or comparisons of large databases. Metabolites specifically mentioned in the text are listed by InChiKeys which will be indexed automatically by Google to link this publication to structures. In addition, all FiehnLib metabolite InChI keys are already indexed by Google through the Fiehnlab library website.

There is a high overlap of metabolites between the FiehnLib GC-quadrupole and GC-TOF mass spectral libraries, but these libraries are not identical and continue to increase in size. First, instrument settings were different and followed vendor-specific parameters, e.g. using a much slower temperature ramp for the slow-scanning quadrupole GC-MS instrument. The most important reason for establishing two different libraries was the large differences in retention times due to the use of different columns and temperature programs. Therefore, the libraries are most useful under the parameters given in the method section. Furthermore, quadrupole instruments are more widely used in laboratories, but TOF analyzers have a higher scan rate and therefore enable better mass spectral deconvolution. TOF and quadrupole MS

spectra were both acquired in unit resolution mode, but spectra differed in ion intensity ratios, specifically for high m/z ions. Spectra were also acquired in different m/z ranges. The quadrupole GC/MS library was acquired more closely to mass spectrometry community standards, starting at low m/z values (e.g. for substructure recognition), whereas the TOF spectra were acquired starting at m/z 85 for direct use in metabolomics of TMS-derivatized samples, to avoid the base peak m/z 73 ions (the TMS fragment itself) as well as m/z 79 tailing from pyridine which otherwise may influence peak finding and correct deconvolution by the LECO software in the metabolomics data sets.

Use of the Agilent GC-Quad and the Leco GC-TOF FiehnLib retention index libraries

Due to the included retention index information, the novel FiehnLib libraries are better suited for unambiguous compound identification than smaller libraries or mass spectral libraries that lack consistent retention information. The FiehnLib libraries can be used in conjunction with commercial systems and support the Agilent ChemStation 'Retention Time Lock' system as well as Leco's ChromaTOF software using 'Retention Index Tables'. Importantly, the libraries also support the open source Binbase automatic annotation system[61] with its class assignments provided by the open source SetupX data base[62] and in principle any other software program that uses retention information and mass spectral matching.

Strategies to align data sets from GC-MS and LC-MS chromatograms have been published for at least 15 years63 as accumulated in a public web resource[64]. Such approaches use chemometric methods to adjust for retention time shifts due to column selectivity changes because of temperature fluctuations, irreversible adsorption, column degradation, machine drifts and other problems. However, all alignment programs are only targeted at aligning 'similar' chromatograms, not very dissimilar samples such as urine and plasma profiles. In addition, it has not been shown that even similar chromatograms (e.g. only using blood plasma) could be aligned over years of analysis, with potential large shifts due to column cuts or column aging.

As the FiehnLib libraries are based on retention indices instead of signal-based chromatogram alignments, metabolites can be identified and compared across highly diverse chromatograms as shown for the BinBase query of 846 different samples covering plants, animals and microbes and selecting studies spanning five years of data acquisition. Such use of metabolomic mass spectral databases has never been shown before in the peer-reviewed literature. Chromatograms of such different matrices, complexity, compound identities and time spans would be impossible to compare by chromatographic alignment tools, without spiked retention index marker compounds and retention index and mass spectral library matching.

As proof-of-concept, the species comparison in figure 5 emphasized how metabolite presence in different species and organs could be comprehensively retrieved, for example for taxonomic or phylogenetic purposes. Conversely, the database can also be investigated for metabolites that are specific biomarkers for a certain disease, for example when querying data sets of 'metabolic phenotypes of different diseases in mammalian blood plasma'. For quantitative purposes, data were normalized to the total sum of all identified metabolites as a surrogate for the absolute concentration per number of cells. Efforts in standardization of metabolomic reports will necessitate that these normalized relative values will be transferred to semi-quantitative assessments of absolute concentrations to compensate for instrument drifts, to better compare quantitative data between studies, between organisms and even between laboratories in Round-Robin-Tests[66] or ring trials.

Conclusion

A mass spectral and retention index library of chemically diverse small molecules was presented to comprehensively and unambiguously identify compounds in metabolite profiling using GC/MS. Based on standard operating procedures and consistent parameters used for over five years, the database was utilized in the UC Davis Genome Center by 104 users (53 UC Davis researchers and 51 external scientists) in 279 metabolomic studies covering 19,032 animal, human, plant and microbial samples. Despite good overlap of the identified metabolites with the KEGG/BioMeta metabolic pathway database, more comprehensive metabolome coverage would require commercial availability of a larger range of chemical standards and substantial resources as well as use of further technologies such as LC-MS/MS. Together, GC/MS and LC-MS/MS libraries eventually may support National Metabolome Repositories similar to NCBI's gene expression[67] and public proteome databases.

Chemicals mentioned in the text with PubChem CIDs and InChIKeys

Methylhydroxylamine Hydrochloride; CID: 521874; XNXVOSBNFZWHBV-UHFFFAOYAC

Pyridine; CID: 1049; JUJWROOIHBZHMG-UHFFFAOYAY

N-Methyl-N-trimethylsilyltrifluoroacetamide; CID: 32510; MSPCIZMDDUQPGJ-UHFFFAOYAZ

Trimethylchlorosilane; CID: 6397; IJOOHPMOJXWVHK-UHFFFAOYAM

Chloroform; CID: 6212; HEDRZPFGACZZDS-UHFFFAOYAG

Helium; CID: 23987; SWQJXJOGLNCZEY-UHFFFAOYAJ

Estrone; CID: 5870; DNXHEGUUPJUMQT-CBZIJGRNBW

(Z)-O-methoxyamino-estrone; PBXZSAMJKKLMEF-DUTNZQCHBS

(E)-O-methoxyamino-estrone; PBXZSAMJKKLMEF-YDKWXADCBY

O-TMS-(Z)-O-methoxyamino-estrone; GIUGKXYPAYLRPS-AFLCNQENBJ

O-TMS-(Z)-O-methoxyamino-estrone; GIUGKXYPAYLRPS-XKNWVNBWBA

Cholic acid; CID: 303; BHQCQFFYRZLCQQ-UHFFFAOYAA

Hexanoyl-glycine; CID: 99463; UPCKIPHSXMXJOX-UHFFFAOYAU

Monooelein; CID: CID 24699; VBICKXHEKHSIBG-UHFFFAOYAT

Inositol; CID: 892; CDAISMWEOUEBRE-GPIVLXJGBG

Alanine; CID: 602; QNAYBMKLOCPYGJ-UHFFFAOYSA-N

Palmitate; CID: 985; IPCSVZSSVZVIGE-UHFFFAOYSA-N

Uric acid; CID: 1175: LEHOTFFKMJEONL-UHFFFAOYSA-N

Creatinine; CID: 588; DDRJAANPRJIHGJ-UHFFFAOYSA-N

Glucose; CID 5793; WQZGKKKJIJFFOK-GASJEMHNSA-N

Glycerol-alpha-phosphate; CID 754; AWUCVROLDVIAJX-UHFFFAOYSA-N

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support for the scientific evaluation of the FiehnLib libraries and the continuing improvement of BinBase is appreciated by funding through NIH (R01 ES013932), NSF (MCB-0820823) and the European Union (FP7 Health-2007-2.1.4.1 project 200327). Data acquisition for the libraries was initiated by a DuPont corporation donation to OF and subsequently funded by contract work with Agilent and Leco corporations. We thank Katherine Gharibian (FiehnLab) for her help with GC-MS standard preparations. We thank Martin Scholz (FiehnLab) for implementing SetupX. We thank Kirsten Skogerson (FiehnLab) for help editing the manuscript. We thank Steve Carano (VWR) and John Nicolas (Fisher Scientific) for help during the substance purchase and ordering phase. We thank Craig Morgan (MDL Elsevier) for matching our structure data against the MDL ACD Available Chemicals Directory. We thank Scott Hutton (ChemNavigator) for matching our structure data against the iResearch Library. We thank ChemAxon for a free academic license of their cheminformatics suite including JChem and Instant-JChem. We thank Steve Stein and the NIST mass spectrometry group for the NIST MS Search DLL source code and the freely available NIST MS Search program. We thank Masanori Arita and Kazuhiro Suwa (metabolome.jp) for providing the NIST mass spectral search in JAVA code under the GNU General Public License.

References

1. Fiehn O. Metabolomics–the link between genotypes and phenotypes. Plant Molecular Biology 2002;48 (1):155–171. [PubMed: 11860207]
2. Horning EC, Horning MG. Metabolic Profiles: Gas-Phase Methods for Analysis of Metabolites. Clin Chem 1971;17(8):802–809. [PubMed: 5105517]
3. Thompson JA, Markey SP. Quantitative metabolic profiling of urinary organic acids by gas chromatography-mass spectrometry. Comparison of isolation methods. Analytical Chemistry 1975;47 (8):1313–1321. [PubMed: 1147254]
4. Halket JM, Waterman D, Przyborowska AM, Patel RKP, Fraser PD, Bramley PM. Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. Journal of Experimental Botany 2005;56(410):219–243. [PubMed: 15618298]
5. Heller SR. Conversational mass spectral retrieval system and its use as an aid in structure determination. Analytical Chemistry 1972;44(12):1951–1961.
6. McLafferty FW, Stauffer DA, Loh SY, Wesdemiotis C. Unknown identification using reference mass spectra. quality evaluation of databases. Journal of the American Society for Mass Spectrometry 1999;10(12):1229–1240. [PubMed: 10584326]
7. Kováts E. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. Helvetica Chimica Acta 1958;41(7):1915–1932.
8. Babushok VI, Linstrom PJ, Reed JJ, Zenkevich IG, Brown RL, Mallard WG, Stein SE. Development of a database of gas chromatographic retention properties of organic compounds. Journal of Chromatography A 2007;1157(1–2):414–421. [PubMed: 17543315]
9. Stein SE, Babushok VI, Brown RL, Linstrom PJ. Estimation of Kovats retention indices using group contributions. Journal of Chemical Information and Modeling 2007;47(3):975–980. [PubMed: 17367127]
10. Lu H, Liang Y, Dunn WB, Shen H, Kell DB. Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS. TrAC Trends in Analytical Chemistry 2008;27(3): 215–227.
11. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, Dormann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D. GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 2005;21(8):1635–1638. [PubMed: 15613389]

12. Horai, H.; Suwa, K.; Arita, M.; Nihei, Y.; Nishioka, T. MassBank - Mass Spectral Database for Metabolome Analysis. [accessed August 2009]. http://www.massbank.jp
13. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G. METLIN - A metabolite mass spectral database. Therapeutic Drug Monitoring 2005;27(6):747–751. [PubMed: 16404815]
14. Staeb JA, Epema OJ, van Duijn P, Steevens J, Klap VA, Freriks IL. Automated storage of gas chromatography–mass spectrometry data in a relational database to facilitate compound screening and identification. Journal of Chromatography A 2002;974(1–2):223–230. [PubMed: 12458939]
15. Adams, RP. Identification of Essential Oil Components by Gas Chromatography/Quadruple Mass Spectrometry. [accessed August 2009]. http://www.allured.com
16. König, Wilfried A.; Joulain, Daniel; Hochmuth, DH. GC/MS Library: Terpenoids and Related Constituents of Essential Oils. [accessed August 2009]. http://www.massfinder.com
17. Fiehn O, Wohlgemuth G, Scholz M. Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata. Lecture Notes in Computer Science 2005;3615:224.
18. Ott MA, Vriend G. Correcting ligands, metabolites, and pathways. BMC bioinformatics 2006;7(1): 517. [PubMed: 17132165]
19. Ott, MA.; Vriend, G. BioMeta database. [accessed August 2009]. http://cheminf.cmbi.ru.nl/biometa/
20. LIPID MAPS . LIPID Metabolites And Pathways Strategy. [accessed August 2009]. http://www.lipidmaps.org/
21. ChemAxon Instant JChem 2.0 - structure database management, search and prediction. [accessed August 2009]. http://www.chemaxon.com
22. Open Babel - library for molecule file conversion and pattern matching. [accessed August 2009]. http://openbabel.sourceforge.net
23. Wagner C, Sefkow M, Kopka J. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochemistry 2003;62 (6):887–900. [PubMed: 12590116]
24. Colebatch G, Desbrosses G, Ott T, Krusell L, Montanari O, Kloska S, Kopka J, Udvardi MK. Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in. The Plant Journal 2004;39:487–512. [PubMed: 15272870]
25. Stein S, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. Journal of the American Society for Mass Spectrometry 1994;5(9):859–66.
26. Arita, M.; Suwa, K. NIST Search for searching MS peaks. [accessed August 2009]. http://www.metabolome.jp/download/ms-software
27. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic acids research 2002;30(1):402. [PubMed: 11752349]
28. Laine RA, Sweeley CC. Analysis of trimethylsilyl O-methyloximes of carbohydrates by combined gas-liquid chromatography-mass spectrometry. Analytical Biochemistry 1971;43(2):533–538. [PubMed: 5141097]
29. Bentley R, Botlock N. A gas chromatographic method for analysis of anomeric carbohydrates and for determination of mutarotation coefficients. Analytical Biochemistry 1967;20(2):312–320. [PubMed: 4292704]
30. Klee, MS.; Wylie, PL.; Quimby, BD.; Blumberg, LM. Method for sample identification using a locked retention time database. [accessed August 2009]. http://www.google.com/patents?hl=en&lr=&vid=USPAT5827946
31. Rostad CEWEP. Kovats and lee retention indices determined by gas chromatography/mass spectrometry for organic compounds of environmental interest. Journal of High Resolution Chromatography 1986;9(6):328–334.
32. Babushok VI, Linstrom PJ. On the relationship between Kovats and Lee retention indices. Chromatographia 2004;60(11–12):725–728.
33. PubChem Power User Gateway (PUG) . [accessed August 2009]. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_pug.pdf

34. Pirok G, Mate N, Varga J, Szegezdi J, Vargyas M, Dorant S, Csizmadia F. Making “real” molecules in virtual space. Journal of Chemical Information and Modeling 2006;46(2):563–568. [PubMed: 16562984]
35. Ott MA, Vriend G. Correcting ligands, metabolites, and pathways. Bmc Bioinformatics 2006;7
36. Flower DR. On the Properties of Bit String-Based Measures of Chemical Similarity. Journal of Chemical Information and Computer Sciences 1998;38(3):379–386.
37. FiehnLib - a mass spectral and retention index library for comprehensive metabolic profiling . [accessed August 2009]. http://fiehnlab.ucdavis.edu/projects/FiehnLib/
38. Demuth W, Karlovits M, Varmuza K. Spectral similarity versus structural similarity: mass spectrometry. Analytica Chimica Acta 2004;516(1–2):75–85.
39. Varmuza K. Recognition of Relationships between Mass Spectral Data and Chemical Structures by Multivariate Data Analysis. ANALYTICAL SCIENCES 2001;17(i467)
40. NIST MS Search Program v2.0. [accessed August 2009]. http://chemdata.nist.gov/
41. SetupX - FiehnLab Open LIMS. [accessed August 2009]. http://fiehnlab.ucdavis.edu:8080/m1/main.jsp
42. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TWM, Fiehn O, Goodacre R, Griffin JL. Proposed minimum reporting standards for chemical analysis. Metabolomics 2007;3 (3):211–221.
43. Brändle, Martin Paul; Zass, E. Chemicals Catalog Databases: An Overview and Evaluation. [accessed August 2009]. http://www.infonortics.com/chemical/ch06/slides/brandle.pdf
44. Irwin JJ, Shoichet BK. ZINC–A Free Database of Commercially Available Compounds for Virtual Screening. Journal of chemical information and modeling 2005;45(1):177. [PubMed: 15667143]
45. eMolecules.com . Free chemical structure search engine. [accessed August 2009]. http://www.emolecules.com/
46. ChemSpider-Database of Chemical Structures and Property Predictions. [accessed August 2009]. http://www.chemspider.com/
47. ChemNavigator - database of commercially accessible screening compounds. [accessed August 2009]. http://www.chemnavigator.com
48. Heller, SR.; Lowry, SR. Library Storage and Retrieval Methods in Infrared Spectroscopy. [accessed August 2009]. http://www.hellers.com/steve/resume/p101.html
49. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TWM, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR. Proposed minimum reporting standards for chemical analysis. Metabolomics 2007;3(3):211–221.
50. Ianni JC, Annamalai V, Phuan PW, Panda M, Kozlowski MC. A priori theoretical prediction of selectivity in asymmetric catalysis: Design of chiral catalysts by using quantum molecular interaction fields. Angewandte Chemie-International Edition 2006;45(33):5502–5505.
51. Avery KA, Mann R, Norton M, Willock DJ. Computer simulation of structural aspects of enantioselective heterogeneous catalysis and the prospects for direct calculation of selectivity. Topics in Catalysis 2003;25(1–4):89–102.
52. Fiehn O, Wohlgemuth G, Scholz M, Kind T, Lee DY, Lu Y, Moon S, Nikolau B. Quality control for plant metabolomics: reporting MSI-compliant studies. Plant Journal 2008;53(4):691. [PubMed: 18269577]
53. Kanani HH, Klapa MI. Data correction strategy for metabolomics analysis using gas chromatography-mass spectrometry. Metabolic Engineering 2007;9(1):39–51. [PubMed: 17052933]
54. Chapman & Hall/CRC Dictionary of Natural Products (DNP) online. [accessed August 2009]. http://dnp.chemnetbase.com/
55. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H. The scaffold tree-Visualization of the scaffold universe by hierarchical scaffold classification. Journal of Chemical Information and Modeling 2007;47(1):47–58. [PubMed: 17238248]
56. Borgelt, C. MoSS - Molecular Substructure Miner. [accessed August 2009]. http://www.borgelt.net/moss.html

57. Borgelt C, Berthold MR, Patterson DE. Molecular fragment mining for drug discovery. Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Proceedings 2005;3571:1002–1013.
58. Medical Subject Headings (MeSH). [accessed August 2009]. http://www.nlm.nih.gov/mesh/MBrowser.html
59. Aoki-Kinoshita KF, Kanehisa M. Gene Annotation and Pathway Mapping in KEGG.
60. Kind T, Scholz M, Fiehn O. How Large Is the Metabolome? A Critical Analysis of Data Exchange Practices in Chemistry. PLoS ONE 2009;4(5)
61. BinBase automatic analysis of chromatograms (Open Source Project). [accessed August 2009]. http://sourceforge.net/projects/binbase/
62. SetupX - study design database for metabolomic projects (Open Source Project). [accessed August 2009]. http://code.google.com/p/setupx/
63. Grung B, Kvalheim OM. Retention time shift adjustments of two-way chromatograms using Bessel's inequality. Analytica Chimica Acta 1995;304(1):57–66.
64. Kind, T. Peak Alignment of LC, GC, MS, NMR data (software collection). [accessed August 2009]. http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Peak_Alignment/
65. Stein SE. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. Journal of the American Society for Mass Spectrometry 1999;10(8):770–781.
66. Allwood JW, Erban A, de Koning S, Dunn WB, Luedemann A, Lommen A, Kay L, Löscher R, Kopka J, Goodacre R. Inter-laboratory reproducibility of fast gas chromatography–electron impact–time of flight mass spectrometry (GC–EI–TOF/MS) based plant metabolomics. Metabolomics :1–18.
67. Gene Expression Omnibus (GEO) database repository of high throughput gene expression data and hybridization arrays, chips, microarrays. [accessed August 2009]. http://www.ncbi.nlm.nih.gov/geo/

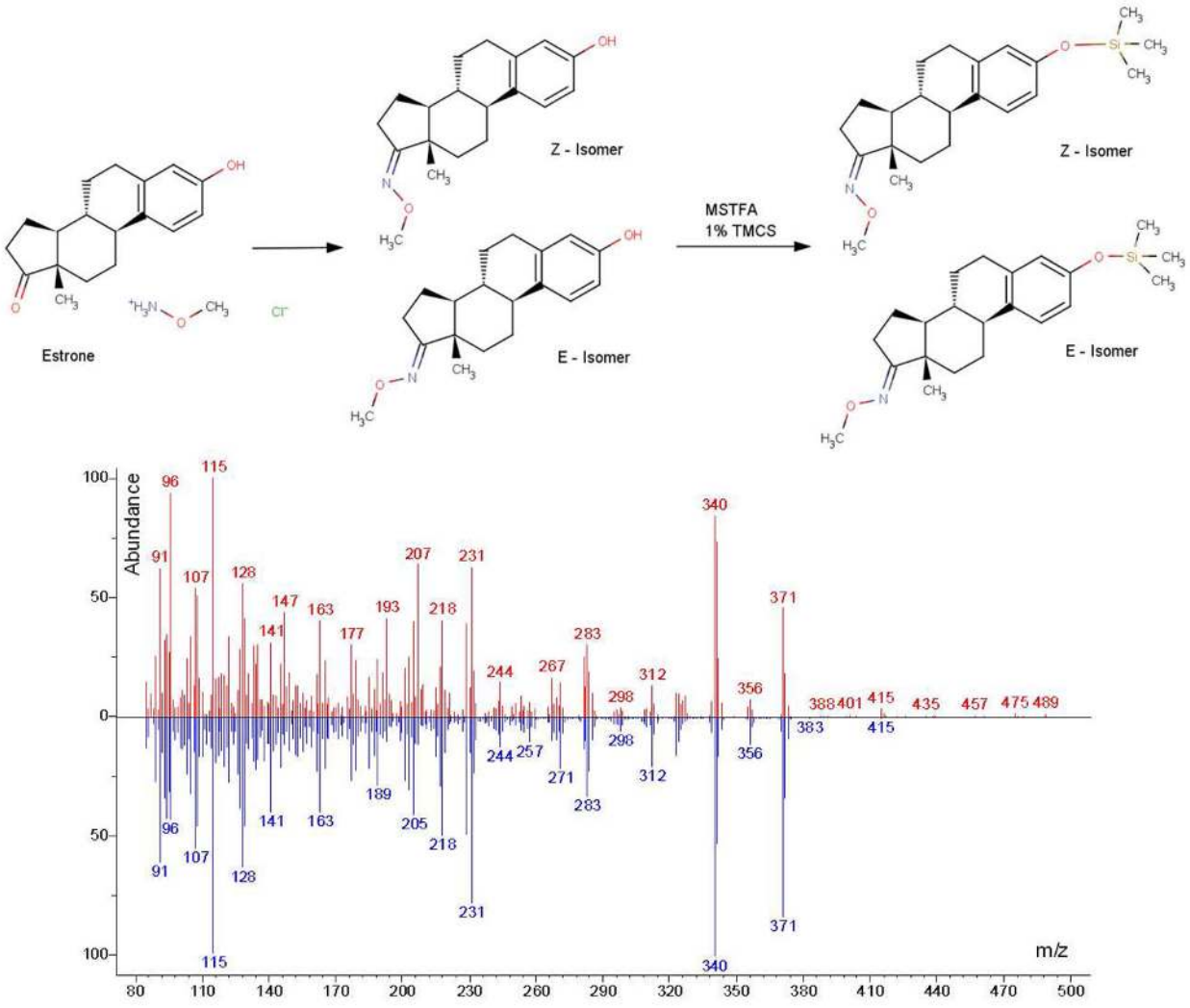

Figure 1.

Derivatization by methoximation and silylation for GC-MS based metabolomics. Upper panel: Derivatization of estrone yielding two isomers (Z- and E-). Lower panel: Comparison of E,Z-trimethylsilyloxy-estrone methoxime mass spectra. Very small mass spectral differences are observed, but peaks are chromatographically resolved with 2s absolute retention time difference.

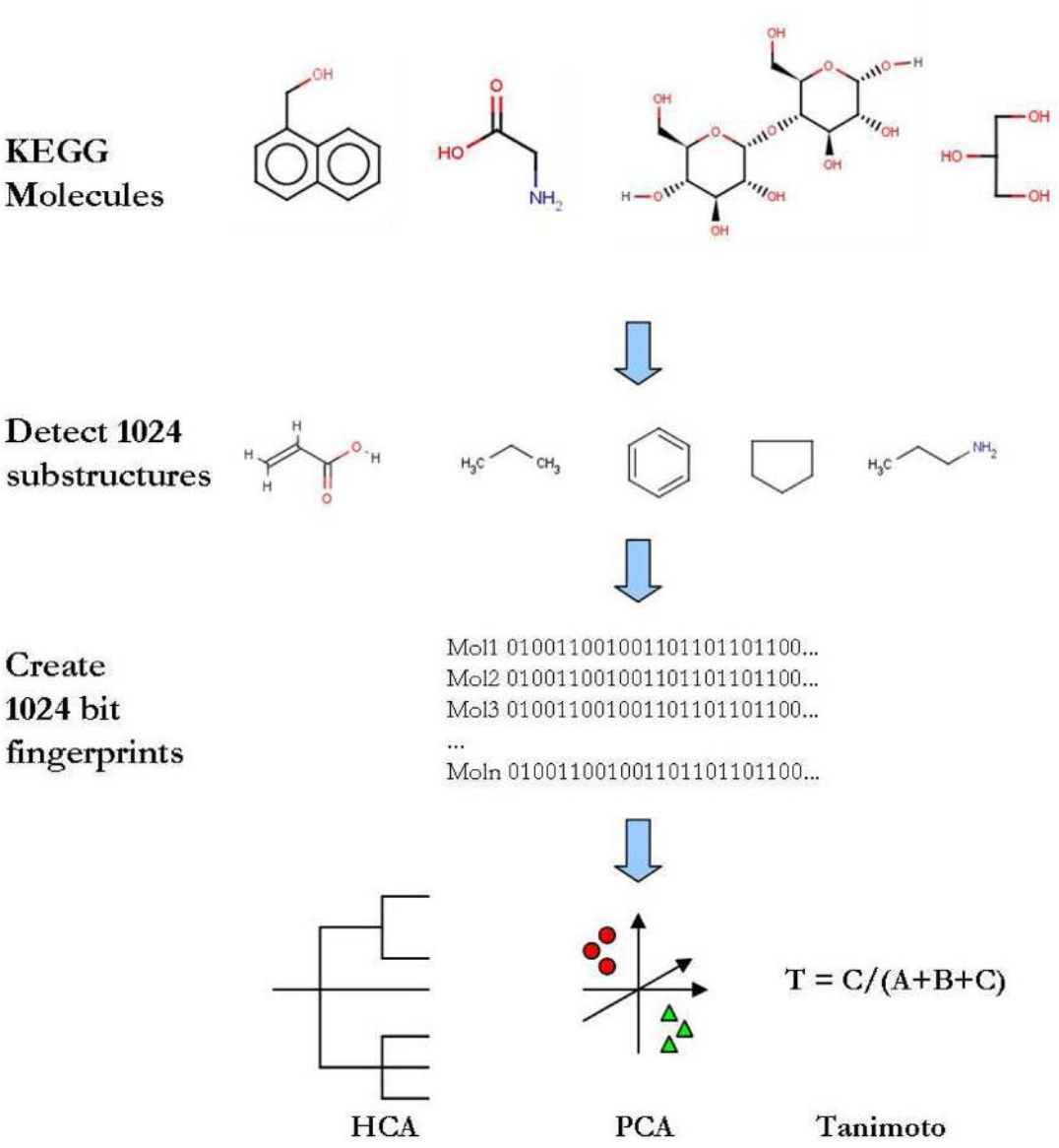

Figure 2.
Calculation of structure overlaps between chemical libraries is performed by creating substructure fingerprints which can be used in principal components analyses (PCA) or Tanimoto similarity calculations.

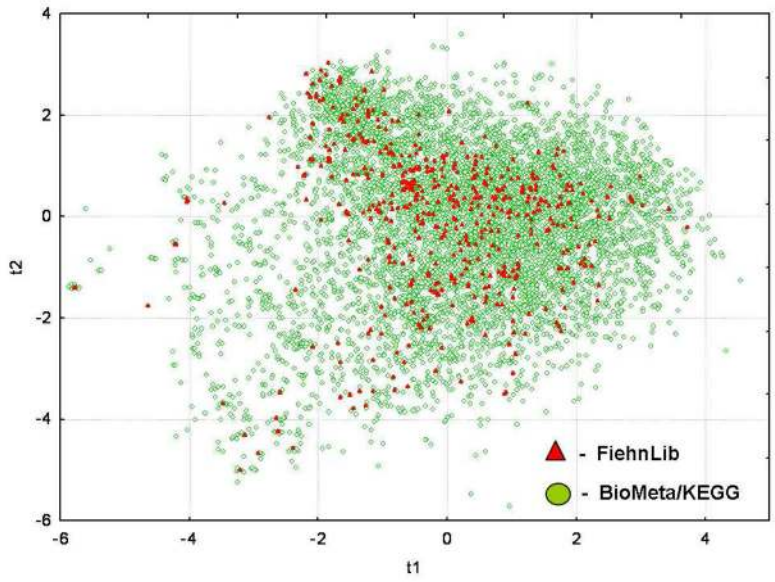

Figure 3.

Principal component analysis (PCA) of chemical hashed fingerprints between the KEGG metabolite database and FiehnLib. The graph demonstrates a high chemical complexity without forming distinct clusters for KEGG metabolites. Fewer compounds are comprised in the FiehnLib retention index/mass spectral library but show a similar structural diversity and overlap density compared to KEGG structures.

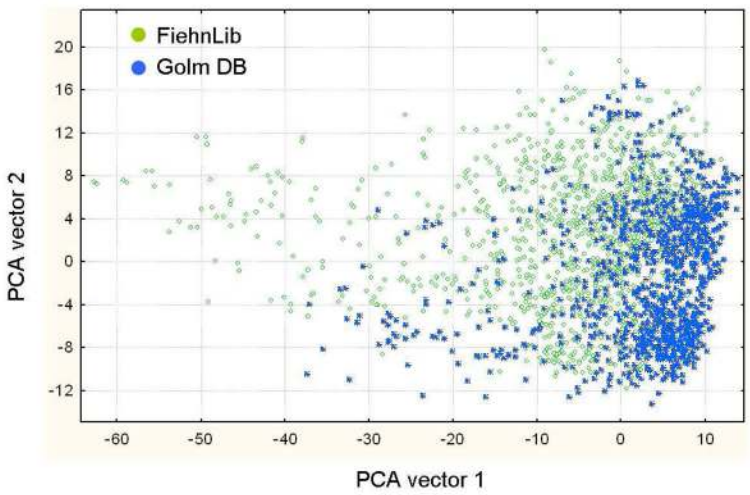

Figure 4.

Overlap of the two mass spectral databases FiehnLib and GolmDB using principal component analysis (PCA). The PCA input matrix was based on mass spectral features. Points which overlap have the same peak similarity; the FiehnLib shows a more diverse distribution of mass spectra, while the Golm DB clusters more strongly.

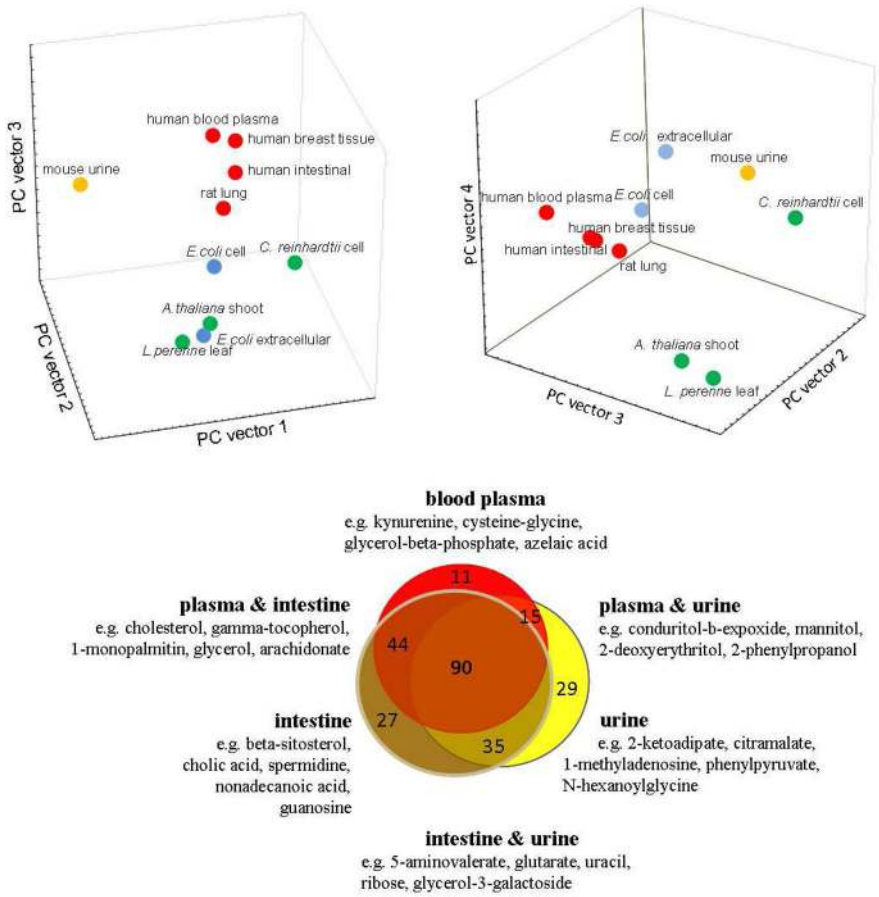

Figure 5.

Results of FiehnLib applications in SetupX/BinBase queries across nine studies (846 samples). Upper panel: Unsupervised Principal Component Analysis graphs on 324 identified unique metabolites. For each species/organ combination, median normalized intensities were calculated including all metabolites that were positively detected in minimum 10% of all samples of a class. Mammalian samples are labeled in red, urine in yellow, E. coli in blue and plants and algae in green.

Lower panel: Venn diagram on a subset of 251 mammalian metabolites identified in plasma, urine and intestinal effluent metabolome. 36% of these metabolites were identified in all three body fluids such as inositol, alanine, palmitate, uric acid, creatinine, glucose and glycerol-alpha-phosphate. 37% of the compounds were only positively detected in two fluids, and 27% were unique for a specific fluid.

Table 1

Details of data acquisition parameters for the FiehnLib GC/MS libraries

	Agilent GC-Quadrupole MS	**Leco GC-Time of Flight MS**
Gas Chromatograph	Agilent 6890GC with Agilent 6890 split/splitless injector	
Mass Spectrometer	Agilent 5973 MSD	Leco Pegasus IV TOF
GC column	10 m Duragard integrated with Agilent 122-5532G DB5-MS 30 m length; 0.25 mm i.d.; 0.25 um film 95% dimethyl/5% diphenyl polysiloxane	10 m guard column, integrated with Restek RTX-5Sil MS 30 m length; 0.25 mm i.d.; 0.25 um film 95% dimethyl/5% diphenyl polysiloxane
GC parameters, injection	1 wash step pre-injection; 4 sample pumps, 10 ul syringe. Injection of 1 ul in sandwich mode with fast plunger speed without viscosity delay or dwell time. Injection at 250°C and split ratio 1:5 to 1:10 with 3-10 ml/min Helium split flow into a Restek 20782 deactivated glass-wool split liner.	
GC parameters, separation	1 ml/min constant flow Helium. Oven ramp 60°C (1 min hold) to 325°C at 10°C/min, 10 min hold before cool-down, 37.5 min run time.	1 ml/min constant flow Helium. Oven ramp 50°C (1 min hold) to 330°C at 20°C/min, 5 min hold before cool-down, 20 min run time.
MS parameters, tuning	Autotune using FC43 (Perfluorotributylamine) with manufacturer-specific tune settings	
MS parameters, data acquisition	Transfer line temperature 290°C Electron impact ionization at 70 eV Filament source temperature 230°C Quadrupole temperature of 150°C Scan range 50-600 u at 2 spectra/s 5.90 min solvent delay time	Transfer line temperature 280°C Electron impact ionization at 70 eV Filament source temperature 250°C TOF at room temperature Scan range 85-500 u at 20 spectra/s 6.50 min solvent delay time
MS parameters, data processing	Peak detection and spectra processing by Agilent ChemStation	Peak detection, deconvolution and spectra processing by Leco ChromaTOF vs. 2.32

Table 2

Contribution (% of total number of functional group hits) of important substructures and classes comprised in FiehnLib and the BioMeta/KEGG database

ID	Functional groups	FiehnLib	BioMeta/KEGG
S12	Alcohol	21.7%	17.8%
SA-6	Carboxylic acid	25.2%	17.0%
S5	Alkenes	7.5%	11.7%
S23	Amines	10.2%	11.6%
S49	Ketones	8.3%	7.2%
SA-5	Nitrogen (n>0) in aromatic 6-ring	4.5%	6.1%
S98	Amides	3.7%	5.7%
S277	Sugar pattern (multiple rings)	3.6%	4.7%
SA-3	Phosphate group containing	3.2%	4.0%
S86	Lactones	1.1%	3.5%
SA-4	Chlorine containing (non salt)	0.0%	2.8%
SA-8	Purines	1.7%	1.7%
SA-7	Carboxyl (acid, ester, salt) with aliphatic carbon chain (n>6)	4.1%	1.5%
S278	Sugar pattern reducing sugars	1.2%	1.3%
S48	Aldehydes	1.5%	1.2%
SA-2	General steroids	1.2%	0.9%
S6	Alkynes	0.0%	0.3%
SA-1	Aromatic steroids	0.5%	0.1%
SA-T	Total number of molecules	701	11280