BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 5 2008, pages 732–737
doi:10.1093/bioinformatics/btn023

Data and text mining

TagFinder for the quantitative analysis of gas chromatography—mass spectrometry (GC-MS)-based metabolite profiling experiments

Alexander Luedemann, Katrin Strassburg, Alexander Erban and Joachim Kopka*

Department Prof. L. Willmitzer, Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, D-14476 Potsdam-Golm, Germany

Received on November 30, 2007; revised on January 9, 2008; accepted on January 12, 2008

Advance Access publication January 19, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Typical GC-MS-based metabolite profiling experiments may comprise hundreds of chromatogram files, which each contain up to 1000 mass spectral tags (MSTs). MSTs are the characteristic patterns of ~25–250 fragment ions and respective isotopomers, which are generated after gas chromatography (GC) by electron impact ionization (EI) of the separated chemical molecules. These fragment ions are subsequently detected by time-of-flight (TOF) mass spectrometry (MS). MSTs of profiling experiments are typically reported as a list of ions, which are characterized by mass, chromatographic retention index (RI) or retention time (RT), and arbitrary abundance. The first two parameters allow the identification, the later the quantification of the represented chemical compounds. Many software tools have been reported for the pre-processing, the so-called curve resolution and deconvolution, of GC-(EI-TOF)-MS files. Pre-processing tools generate numerical data matrices, which contain all aligned MSTs and samples of an experiment. This process, however, is error prone mainly due to (i) the imprecise RI or RT alignment of MSTs and (ii) the high complexity of biological samples. This complexity causes co-elution of compounds and as a consequence non-selective, in other words impure MSTs. The selection and validation of optimal fragment ions for the specific and selective quantification of simultaneously eluting compounds is, therefore, mandatory. Currently validation is performed in most laboratories under human supervision. So far no software tool supports the non-targeted and user-independent quality assessment of the data matrices prior to statistical analysis. TagFinder may fill this gap.

Strategy: TagFinder facilitates the analysis of all fragment ions, which are observed in GC-(EI-TOF)-MS profiling experiments. The non-targeted approach allows the discovery of novel and unexpected compounds. In addition, mass isotopomer resolution is maintained by TagFinder processing. This feature is essential for metabolic flux analyses and highly useful, but not required for metabolite profiling. Whenever possible, TagFinder gives precedence to chemical means of standardization, for example, the use of internal reference compounds for retention time calibration or quantitative standardization. In addition, external standardization is supported for both compound identification and calibration. The workflow of TagFinder comprises, (i) the import of fragment ion data, namely mass, time and arbitrary abundance (intensity), from a chromatography file interchange format or from peak lists provided by other chromatogram pre-processing software, (ii) the annotation of sample information and grouping of samples into classes, (iii) the RI calculation, (iv) the binning of observed fragment ions of equal mass from different chromatograms into RI windows, (v) the combination of these bins, so-called mass tags, into time groups of co-eluting fragment ions, (vi) the test of time groups for intensity correlated mass tags, (vii) the data matrix generation and (viii) the extraction of selective mass tags supported by compound identification. Thus, TagFinder supports both non-targeted fingerprinting analyses and metabolite targeted profiling.

Availability: Exemplary TagFinder workspaces and test data sets are made available upon request to the contact authors. TagFinder is made freely available for academic use from http://www-en.mpimp-golm.mpg.de/03-research/researchGroups/01-dept1/Root_Metabolism/smp/TagFinder/index.html

Contact: Kopka@mpimp-golm.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online and within the TagFinder download from the above URL.

1 INTRODUCTION

Gas chromatography hyphenated to mass spectrometry (GC-MS) is one of the most versatile and widely applied technology platforms in modern metabolomic and fluxomics studies. Metabolic phenotyping has become an integral part of molecular physiology and functional genomics (e.g. Bino *et al.*, 2004; Nicholson *et al.*, 1999; Nielsen and Oliver, 2005; Stephanopoulos *et al.*, 2004; Sumner *et al.*, 2003; Trethewey *et al.*, 1999). As a consequence initiatives for the standardization of high-throughput metabolite analyses have been initiated (Castle *et al.*, 2006; Fiehn *et al.*, 2006; Jenkins *et al.*, 2004; Lindon *et al.*, 2005; Spasić *et al.*, 2006). Guidelines are now available from 10 articles of a recent issue of the Metabolomics journal (Vol. 3, 2007) which comprehensively cover aspects of metabolomics standardization (e.g. Fiehn *et al.*, 2007; Sumner *et al.*, 2007).

*To whom correspondence should be addressed.

However, contrary to the efforts of unifying the standards for data retrieval, mining and interpretation, highly diverse and in part specialized software solutions for the GC-MS data pre-processing have been published. The pre-processing for metabolomic studies typically deals with the seemingly simple task of transforming chemical compound information and respective peak lists from series of chromatography data files into numeric data matrices, which are then amenable to statistical analysis. Both automated peak extraction, and the automated deconvolution of mass spectra allow the comprehensive and non-biased analysis of GC-MS experiments. Previous knowledge about all potential compounds, which may occur in a sample, is not required. In addition, deconvolution reconstructs mass spectra and, thus, provides in principle relevant mass spectra for compound identification. Thus, deconvolution has been an intensely explored topic. In the following, we will shortly summarize basic categories of pre-processing.

The conventional approach of mass spectral deconvolution is based on information present within single chromatograms. Just to mention one early approach, Pool and co-workers developed a backfolding procedure for the mathematical enhancement of GC-MS-based chromatographic curve resolution (Pool *et al*., 1996, 1997a, b). In contrast, multivariate curve resolution (MCR) and its successor the hierarchical multivariate curve resolution (HDA) may represent the most advanced pre-processing tool. Information from multiple aligned chromatograms is utilized for joined deconvolution (Jonsson *et al*., 2004, 2005, 2006). The benefit of this multi-chromatogram procedure is the separation of co-eluting MSTs based on the independence of concentration changes of different compounds in many samples. Even mixtures of exactly co-eluting ambient ^{12}C- and fully stable isotope labelled ^{13}C-mass isotopomers can be resolved using MCR (Kopka J, personal communication). However, MCR is highly sensitive to the selection and number of co-analysed chromatogram files and does not allow targeted retrieval of selected fragment ions or extraction of mass isotopomer distributions for flux analysis.

Perhaps, the most widely spread tool may be the mass spectral deconvolution and identification system (AMDIS; http://chemdata.nist.gov/mass-spc/amdis/overview.html; Halket *et al*., 1999; Stein, 1999), which is provided with the standard mass spectral search and comparison software NIST05 (National Institute of Standards and Technology, Gaithersburg, MD, USA; http://www.nist.gov/srd/mslist.htm). AMDIS was initially designed for purely qualitative analysis, but now also extracts quantitative information, such as the base peak intensity and an estimation of the total intensity of each deconvoluted MST. The main commercial competitor of the universally applicable AMDIS tool is the ChromaTof software (LECO, St. Joseph, MI, USA; http://www.leco.org/), which is exclusive for the GC-EI-TOF-MS and two-dimensional GCxGC-EI-TOF-MS instruments of the vendor. ChromaTof software is designed for in-line acquisition and analysis of GC-TOF-MS chromatograms. The software is customized for the high frequency of mass spectral acquisition (10–500 scans s^{-1}) and the resulting large file sizes, which are generated by fast scanning TOF technology (e.g. Dalluge *et al*., 2002a, b; Vreuls *et al*., 1999). Automated mass spectral deconvolution appears to be highly successful for compound discovery, but comes at the price of software errors, such as partially deconvoluted MSTs, mixed or in other words chimeric MSTs, occurrence of artificial MSTs due to electronic noise, and erroneous MST duplications.

For the improvement of data analysis and retrieval from metabolite profiling experiments, software projects were initiated in academia, which were based on the comprehensive extraction of mass selective peak apex intensities. This approach is computationally less demanding compared to the, traditionally preferred, extraction of peak areas. A typical example of these software tools is MetAlign (http://www.pri.wur.nl/UK/products/MetAlign/). MetAlign was initially commercialized and is now free for academic use. It was successfully targeted at supporting LC-MS analyses and includes highly parameterized smoothing, baseline correction and statistical chromatographic alignment options. These options were recently extended by a mass alignment procedure for the accommodation of high mass-accuracy instruments (America *et al*., 2006; Bino *et al*., 2005; De Vos *et al*., 2007; Keurentjes *et al*., 2006; Vorst *et al*., 2005). MetAlign is applicable to GC-EI-TOF-MS analysis (Tikunov *et al*., 2005), but appears not to perform deconvolution, in other words the combination of extracted mass tags into full MSTs. Further software tools allow both non-targeted as well as compound-targeted GC-MS analysis, for example the tools, XCMS (Smith *et al*., 2006), MathDAMP (Baran *et al*., 2006), MetaQuant (Bunk *et al*., 2006), the MSFACTs (Duran *et al*., 2003; http://www.noble.org/PlantBio/MS/MSFACTs/MSFACTs.html) and the respective refinement, MET-IDEA Broeckling *et al*., 2006; (http://www.noble.org/Plantbio/MS/MET-IDEA/index.html) or the progressive peak clustering approach (De Souza *et al*., 2006). First attempts have also been made at compound-targeted processing of GCxGC-EI-TOF-MS files (e.g. Sinha *et al*., 2004). One unique tool, the BinBase (Fiehn *et al*., 2005; http://fiehnlab.ucdavis.edu/projects/binbase_setupx/), combines GC-EI-TOF-MS data analysis with a database application, which collects and archives extracted MSTs. Other MST and mass spectral libraries were made publicly available (e.g. Kopka *et al*., 2005; Schauer *et al*., 2005; Wagner *et al*., 2003; http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html) or may be commercially obtained, such as the Wiley library (http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470047860.html) or the NIST05 collection, mentioned above (Ausloos *et al*., 1999; Halket *et al*., 2005; http://chemdata.nist.gov/mass-spc/Srch_v1.7/index.html).

In summary, a rich and diverse resource of freely or commercially available tools exists for the pre-processing of GC-MS chromatography data. Each tool has a specific set of parameter settings, which needs to be optimized by human intervention. All tools will in general provide correct and repeatable processing results. But results from different tools are currently difficult to compare. Moreover, most tools do not retrieve mass isotopomer distributions for flux analysis. TagFinder represents the solution to data retrieval for flux studies and allows the comparative analysis of pre-processing procedures. In the following, we will shortly describe the generalized workflow of a TagFinder supported analysis of a GC-(EI-TOF)-MS profiling experiment.

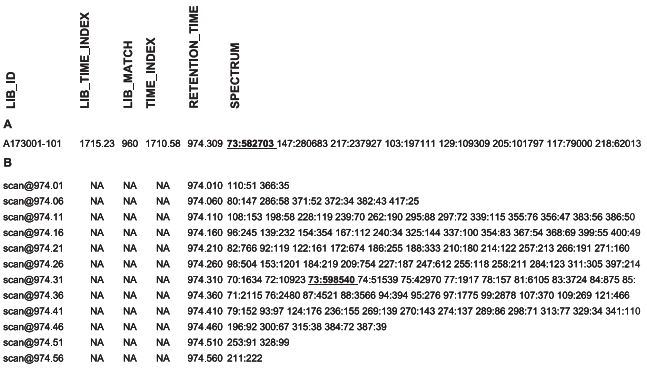
2 WORKFLOW

2.1 Data import

TagFinder software (Supplementary file 1) is a single user application for personal computer systems based on the Java™ programming language and Java™ runtime environment 1.5 or higher (http://www.java.com/de/download/). TagFinder is based on workspaces, which define the mass range and the decimal precision of the RI system. The RI precision is user-defined and should be based on the speed and rate of data acquisition of the GC-MS experiment under investigation. The recommended data import uses the commonly accepted chromatography interchange format NetCDF, which can be exported from almost any vendors' GC-MS acquisition software. TagFinder may utilize non-processed exports. However, the export of baseline corrected and smoothed NetCDF files is recommended. Smoothing and baseline correction are as a rule best performed by vendor and system specific software applications, because different GC-MS technologies may require specific parameter settings and algorithms, which should be optimized by each vendor. An interface for the use of MetAlign for chromatography data pre-processing is available upon request.

TagFinder generates a peak list which corresponds to each NetCDF file. The name of the peak list file will be identical to the name of the processed NetCDF file and is used for subsequent unambiguous annotation of sample information. The peak list files for import into TagFinder software contain mass fragments, specifically the chromatographic peak apices which are linked to mass, RT and peak height information. The mass, retention- and intensity-ranges may be customized prior to data import. TagFinder provides an additional import option for mass spectral deconvolutions and matching results from the ChromaTof software. In detail, the name of the best mass spectral hit, the respective matching factor, the expected RI and the measured RT are imported. If the RI is calculated through an external software, the pre-calculated RI may be imported rather than RT. Two typical structures of TagFinder import files are shown (Fig. 1). These examples demonstrate a typical ChromaTof deconvolution of a single chromatographic peak (Fig. 1A) and the respective result of the same peak after NetCDF file processing and peak apex extraction (Fig. 1B).

Our reference data set for performance testing comprises 32 NetCDF files of approximately 158.100 KB each, generated on a Pegasus III system (LECO, St. Joseph, MI, USA), which are reduced to 57.000–76.000 KB after baseline correction by the ChromaTof software (Version 1, 2002, Pegasus driver 1.61). ChromaTof deconvolution and processing was 20 min per file and generated 600–1.400 KB peak list files in tab delimited text format (cf. Supplementary file 2). TagFinder required ~120 s per baseline corrected NetCDF file for peak list processing and generated 450–850 KB tab delimited text files (cf. Supplementary file 3). These performance data were estimated using a 2.26 GHz, 2.00 GB RAM single Intel Pentium M processor laptop computer with a Microsoft Windows XP Professional operating system (Version 2002, service pack 2). The subsequent performance data were generated using either the ChromaTof processed or the TagFinder processed compendium of reference peak lists. The minimum imported intensity



LIB_ID	LIB_TIME_INDEX	LIB_MATCH	TIME_INDEX	RETENTION_TIME	SPECTRUM
A					
A173001-101	1715.23	960	1710.58	974.309	**73:582703** 147:280683 217:237927 103:197111 129:109309 205:101797 117:79000 218:62013
B					
scan@974.01	NA	NA	NA	974.010	110:51 366:35
scan@974.06	NA	NA	NA	974.060	80:147 286:58 371:52 372:34 382:43 417:25
scan@974.11	NA	NA	NA	974.110	108:153 198:58 228:119 239:70 262:190 295:88 297:72 339:115 355:76 356:47 383:56 386:50
scan@974.16	NA	NA	NA	974.160	98:245 139:232 154:354 167:112 240:34 325:144 337:100 354:83 367:54 368:69 399:55 400:49
scan@974.21	NA	NA	NA	974.210	82:766 92:119 122:161 172:674 186:255 188:333 210:180 214:122 257:213 266:191 271:160
scan@974.26	NA	NA	NA	974.260	98:504 153:1201 184:219 209:754 227:187 247:612 255:118 258:211 284:123 311:305 397:214
scan@974.31	NA	NA	NA	974.310	70:1634 72:10923 **73:598540** 74:51539 75:42970 77:1917 78:157 81:6105 83:3724 84:875 85:
scan@974.36	NA	NA	NA	974.360	71:2115 76:2480 87:4521 88:3566 94:394 95:276 97:1775 99:2878 107:370 109:269 121:466
scan@974.41	NA	NA	NA	974.410	79:152 93:97 124:176 236:155 269:139 270:143 274:137 289:86 298:71 313:77 329:34 341:110
scan@974.46	NA	NA	NA	974.460	196:92 300:67 315:38 384:72 387:39
scan@974.51	NA	NA	NA	974.510	253:91 328:99
scan@974.56	NA	NA	NA	974.560	211:222

Fig. 1. *The data import structure of TagFinder*. The elution time window (±0.3 s) of the compound ribitol (5TMS) from a single chromatogram is shown. Pre-processing was performed using either ChromaTof deconvolution (**A**) or the NetCDF file pre-processing implemented in TagFinder (**B**). ChromaTof software performs mass spectral deconvolution of MSTs, mass spectral matching and RI calculation, whereas NetCDF processing only extracts mass tags. The matched name (LIB_ID) is an analyte identifier taken from GMD (http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html). The mass spectral display is truncated and represents non-normalized abundances. The respective values of m/z 73 are underlined. Typical peak lists are shown in supplementary files 2 and 3.

was set to 25 arbitrary abundance units. Import generates a single file from all initial peak lists. Our TagFinder processed reference compendium of 32 chromatograms generated a 1.36 GB TagFinder file within ~167.1 s, whereas the ChromaTof pre-processing reduced the Tagfinder file to 0.1 GB, which is imported within a few seconds. Tagfinder files, which are substantially larger than 2.04 GB, cannot be handled due to limitations of the operating system and JAVA™ run time environment. However, the intensity threshold for data import can be optimized to use the full TagFinder file size. For example, the above 1.36 GB TagFinder file can be reduced to 0.84 GB or 0.58 GB with a 50 or 100 arbitrary abundance unit cut-off, respectively. Thus larger sets of chromatograms can be simultaneously analysed. The advised intensity limit for the processing of up to 250 GC-TOF-MS chromatogram sets using TagFinder is approx. 100–250 arbitrary units.

2.2 Annotation of sample classes

The annotation of sample classes is performed prior to numerical data processing. Sample classifications may be used for supervised data processing methods and, therefore, respective information is established within TagFinder. In order to avoid erroneous annotation, the complete sample information is provided in a sample header section of the final data matrix (cf. Section 2.7.). Sample information may be entered manually or can be imported from a tab delimited sample annotation file. The required table must have a column labelled 'RAWNAME', which contains the names of all imported data files (cf. Section 2.1.; naming of the peak list files). All other sample classifications and information can be freely defined. A column labelled 'AMOUNT' is suggested, which should contain information on the sample amount or volume

required for subsequent data normalization after TagFinder processing.

2.3 RI calculation

RI calculation is performed using the retention time of internal reference substances, such as *n*-alkanes or fatty acid methyl esters. RI definitions are made by the user. TagFinder searches for the retention times of added internal reference substances using specific, unique, user-defined fragment ions. Single fragment ions or full and partial mass spectra can be used for the queries, which are restricted to customized retention time windows. Ambiguous results are solved by adjustment of the retention time window, user intervention or automated selection of the highest intensity fragment ion within the selected RT window. A result file is generated and stored, which contains the RI definitions, and the corresponding retention times of each NetCDF file. We recommend storage of this retention time file and of the sample annotation file for documentation within the respective TagFinder workspace. The retention time file is used for automated RI calculation using linear interpolation between the retention time anchors (van den Dool and Kratz, 1963). The user supported generation of the retention time file from our TagFinder pre-processed example requires 10–20 min. The RI calculation of this exemplary data set is completed within ~114.1 s. TagFinder pre-processing requires search for single or pairs of unique mass fragments, whereas ChromaTof pre-processing provides full mass spectra, which can be queried by generic mass spectra of the compound class used to calibrate chromatographic retention.

2.4 Mass tag generation

Mass fragments of all chromatography files from an experiment are sorted by mass and calculated RI. Within this chromatography sorted array RI gaps are scanned, which separate mass fragments of equal mass. This process aligns and bins mass fragments of equal mass across all files of an experiment allowing for the technical variability of RI determination. The bins are in the following called mass tags. Mass tags receive the properties minimum, maximum RIs, RI width and median RI and an average intensity. The main selectable parameters for the gap finding process are, (i) the scanning distance between mass tags and (ii) the minimum width of mass tags. In addition the minimum abundance, maximum width of mass tags, number of mass occurrences among all files or within groups of replicate samples can be set.

Exclusions-masses, as well as mass- and RI-windows allow test runs for selected chromatographic regions and mass ranges. Trials prior to the final generation of a data matrix should be performed. The restriction to a narrow RI window is recommended for parameter optimization. Each GC-MS experiment differs with respect to the absolute concentrations and concentration range of hundreds of compounds. Therefore, previously optimized parameter settings should be critically revised and adapted for each new GC-TOF-MS profiling experiment.

In the course of mass tag generation TagFinder allows the selection of either, the maximum, the average or the sum of intensities for aggregation of mass fragments with equal mass within the same peak list file. This aggregation procedure is necessary, because a single peak list and the original chromatogram may contain more than one mass fragment of equal mass within the generated RI window of the mass tag. This observation may be unexpected, but these options had to be implemented to solve the typical deconvolution errors of AMDIS or ChromaTof software. In addition, multiple peak apices may occur, if chromatogram files with a high acquisition rate or technical noise can only be insufficiently smoothed.

2.5 Time group combination

Mass tags are grouped by TagFinder into so-called time groups using the overlap of RI windows. Mass spectra of each time group are then reconstructed for mass spectral matching using—in a first simplified approach—the average intensities of mass tags from all chromatograms of an experiment. All mass tags exhibiting similar RI median are grouped. For this purpose the mass tags are first arrayed according to ascending RI median. Then steep, stepwise increases of the RI median are detected in order to separate consecutive time groups (Fig. 2). The resolving criteria for time groups are the lack of overlap between the median RI of the preceding mass tag compared to the minimum RI of the following.

2.6 Time group clustering

Time groups of complex metabolite profiles typically contain mass tags of multiple co-eluting compounds. In addition, mass tags which are non-specific may occur. Non-specific mass tags result from compounds with similar chemical moieties, which may cause fragments of identical mass. Pearson or Spearman correlation is applied to the intensity vectors of mass tags of the same time group. This procedure finds correlated clusters of mass tags. Significance and the coefficient of correlation can be set to vary the stringency of time group clustering. The clustering approach uses the observation of most mass spectrometric devices, namely that the intensity ratios of mass

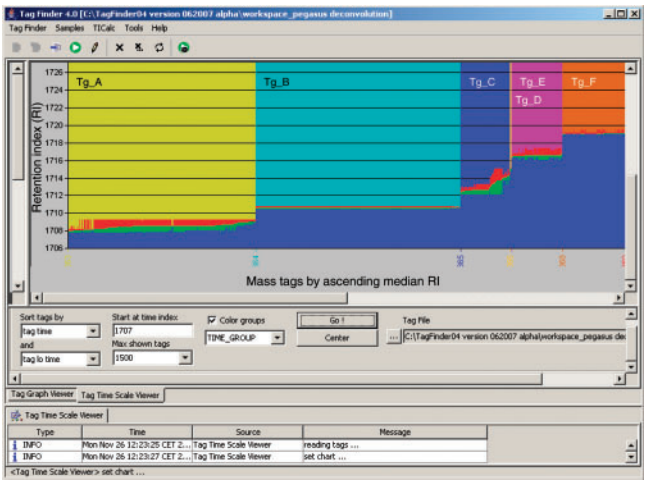


Fig. 2. *Time grouping and clustering*. The screen shot shows the alignment of mass tags by ascending median RI. Green and red bars indicate minimum and maximum RI of each mass tag, the red to green boundary represents median RI. Coloured underlay demonstrates the grouping into time groups (Tg) A–F.

fragments representing a single compound are constant and concentration independent within the range of quantification.

Like the reconstruction of mass spectra from time groups the constituent clusters are transformed into mass spectra using average intensities. The size of clusters can be restricted to a minimum number of mass tags per cluster. Mass tags, which are not correlated according to the significance and correlation coefficient thresholds, are maintained within the data matrix for non-biased fingerprinting analysis, but may also be discarded from further analysis.

2.7 Matrix generation

The data matrix generation is performed after time grouping and time group clustering. The matrix contains all initial intensity data, namely the non-normalized abundance data of each observed mass. Sample information is attached to the header section (cf. Section 2.2). Time group and cluster assignments, mass information, RI data and mass spectral matching results, as far as imported from external processing, are attached to the mass tag row information (cf. Supplementary File 4). The matrix is generated as a tab-delimited text file, which can be imported into statistical tools for data normalization, transformation and statistical analysis. We recommend the TM4 multi-experiment viewer (Saeed *et al*., 2003; Saeed *et al*., 2006) for a first visual assessment.

2.8 Retrieval of selective mass tags

We use libraries of mass spectra and RI of authenticated reference compounds (Kopka *et al*., 2005; Schauer *et al*., 2005; Wagner *et al*., 2003) for compound identification through the matching of reconstructed mass spectra representing time groups or respective clusters within the exported data matrix (Fig. 3).

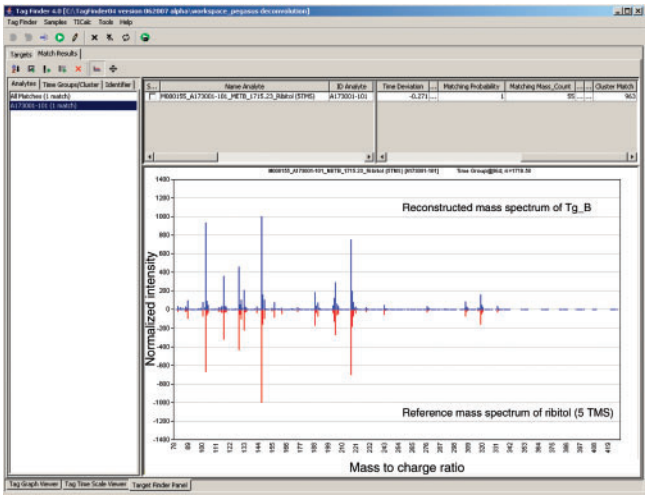


Fig. 3. *The target finding window*. The inset figure shows the head to tail matching of a reference mass spectrum obtained from an authenticated ribitol preparation, bottom (red) and the reconstructed mass spectrum of the respective time group, top (blue). Deviation of retention index (%), number of matching mass tags and the match value are shown.

In addition, the matching can be performed either by using the mass spectral reconstruction from all chromatograms of an experiment or by choice of selected chromatograms, which should contain a known set of targeted reference compounds. We advise to use such reference mixtures for external standardization of each profiling experiment. These reference mixtures, if integrated into the final data matrix, solve ambiguities of compound identification, especially of those chemical isomers, which can only be distinguished by RI and not by mass spectral criteria.

TagFinder extracts all mass tags of time groups based on compound identification by mass spectral matching within expected RI windows. For an improved selectivity, the identified constituent cluster may be obtained and the retrieval of predefined mass fragments is enabled. Single or small sets of selective predefined masses may be useful for comparison of results between data matrices of multiple profiling experiments. The retrieval of targeted fragment masses is also highly useful for mass isotopomer ratio profiling using GC-TOF-MS (Birkemeyer *et al*., 2005) or for flux analysis as demonstrated by Huege *et al*. (2007), which necessitates the retrieval of mass isotopomer distributions.

3 CONCLUSION

In conclusion, we offer a software tool for the alignment of large GC-MS-based metabolite profiling experiments into statistically accessible data matrices. The matrix generation is directed by co-analysis of RI marker substances within each chromatogram and the simultaneous in-parallel analysis of mixtures of reference compounds is recommended. In addition, we offer automated extraction of quantitative data from predefined mass fragments, time groups of mass fragments or clusters of intensity-correlated mass fragments. This extraction of quantitative data is supported by mass spectral matching to reference mass spectra within preset RI windows as are provided by reference libraries (e.g. Kopka *et al*., 2005; Schauer *et al*., 2005). The data matrix generation and matching procedures allow both automation and user intervention for parameter optimization. Thus we present, what we think is the ideal tool for modern metabolomics and fluxomics studies. TagFinder supports non-biased metabolomic fingerprinting, footprinting and profiling experiments (e.g. Kopka *et al*., 2004, 2006a, b) and, moreover, the metabolite targeted analysis of changes in both metabolite pools and flux.

ACKNOWLEDGEMENTS

The authors acknowledge the long standing support and encouragement by Prof. L. Willmitzer, Max Planck Institute of Molecular Plant Physiology (MPI-MP), Am Muehlenberg 1, D-14476 Potsdam-Golm, Germany. We thank Prof. J. Selbig, University of Potsdam, D-14476 Potsdam-Golm, Germany and Dr D. Walther, MPI-MP, for fruitful discussions. This work was supported by the Max Planck Society, the Bundesministerium für Bildung und Forschung (BMBF), grant PTJ-BIO/0312854 and the European META-PHOR project, FOOD-CT-2006-036220.

Conflict of Interest: none declared.

REFERENCES

America,A.H.P. *et al.* (2006) Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional liquid chromatography mass spectrometry. *Proteomics*, **6**, 641–653.

Ausloos,P. *et al.* (1999) The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.*, **10**, 287–299.

Baran,R. *et al.* (2006) MathDAMP: a package for differential analysis of metabolite profiles. *BMC Bioinformatics*, **7**, 530.

Bino,R.J. *et al.* (2004) Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.*, **9**, 418–425.

Bino,R.J. *et al.* (2005) The light-hyperresponsive high pigment-2 mutation of tomato: alterations in the fruit metabolome. *New Phytol.*, **166**, 427–438.

Birkemeyer,C. *et al.* (2005) Metabolome analysis: the potential of *in vivo* labeling with stable isotopes for metabolite profiling. *Trends Biotechnol.*, **23**, 28–33.

Broeckling,C.D. *et al.* (2006) MET-IDEA: Data extraction tool for mass spectrometry-based metabolomics. *Anal. Chem.*, **78**, 4334–4341.

Bunk,B. *et al.* (2006) MetaQuant: a tool for the automatic quantification of GC/MS-based metabolome data. *Bioinformatics*, **22**, 2962–2965.

Castle,A.L. *et al.* (2006) Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. *Briefings Bioinformatics*, **7**, 159–165.

Dalluge,J. *et al.* (2002a) Optimization and characterization of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection (GC × GC-TOF MS). *J. Sep. Sci.*, **25**, 201–214.

Dalluge,J. *et al.* (2002b) Comprehensive two-dimensional gas chromatograpy with time-of-flight mass spectrometric detection applied to the determination of pesticides in food extracts. *J. Chromatogr. A*, **965**, 207–217.

De Souza,D.P. *et al.* (2006) Progressive peak clustering in GC-MS metabolomic experiments applied to Leishmania parasites. *Bioinformatics*, **22**, 1391–1396.

De Vos,R.C.H. *et al.* (2007) Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols*, **2**, 778–791.

Duran,A.L. *et al.* (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics*, **19**, 2283–2293.

Fiehn,O. *et al.* (2005) Automatic annotation of metabolomic mass spectra by integrating experimental metadata. *Proc. Lect. Notes Bioinformatics*, **3615**, 224–239.

Fiehn,O. *et al.* (2006) Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. *Omics- J. Intergrat. Biol.*, **10**, 158–163.

Fiehn,O. *et al.* (2007) The metabolomics standards initiative (MSI). *Metabolomics*, **3**, 175–178.

Halket,J.M. *et al.* (1999) Deconvolution gas chromatography mass spectrometry of urinary organic acids – potential for pattern recognition and automated identification of metabolic disorders. *Rapid Commun. Mass Spectrom.*, **13**, 279–284.

Halket,J.M. *et al.* (2005) Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *J. Exp. Bot.*, **56**, 219–243.

Huege,J. *et al.* (2007) GC-EI-TOF-MS analysis of in vivo carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after ($^{13}CO_2$)-labelling. *Phytochemistry*, **68**, 2258–2272.

Jenkins,H. *et al.* (2004) A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.*, **22**, 1601–1606.

Jonsson,P. *et al.* (2004) A strategy for identifying differences in large series of metabolomic samples analysed by GC/MS. *Anal. Chem.*, **76**, 1738–1745.

Jonsson,P. *et al.* (2005) High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Anal. Chem.*, **77**, 5635–5642.

Jonsson,P. *et al.* (2006) Predictive metabolite profiling applying hierarchical multivariate curve resolution to GC-MS data – a potential tool for multi-parametric diagnosis. *J. Proteome Res.*, **5**, 1407–1414.

Keurentjes,J.J.B. *et al.* (2006) The genetics of plant metabolism. *Nat. Genetics*, **38**, 842–849.

Kopka,J. *et al.* (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.*, **5**, 109–117.

Kopka,J. *et al.* (2005) GMD@CSBDB: The Golm metabolome database. *Bioinformatics*, **21**, 1635–1638.

Kopka,J. (2006a) Current challenges and developments in GC-MS based metabolite profiling technology. *J. Biotechnol.*, **124**, 312–322.

Kopka,J. (2006b) Gas chromatography mass spectrometry. In: Nagata,T., Lörz,H., Widholm,J.M. (eds.) *Biotechnology in agriculture and forestry* Vol. 57: Saito,K, Dixon,RA, Willmitzer,L (eds) Plant metabolomics. Springer-Verlag, Berlin Heidelberg New York, pp. 3–20.

Lindon,J.C. *et al.* (2005) The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics*, **6**, 691–699.

Nicholson,J.K. *et al.* (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, **29**, 1181–1189.

Nielsen,J. and Oliver,S. (2005) The next wave in metabolome analysis. *Trends Biotechnol.*, **23**, 544–546.

Pool,W.G. *et al.* (1996) Backfolding applied to differential gas chromatography/mass spectrometry as a mathematical enhancement of chromatographic Resolution. *J. Mass Spectrom.*, **31**, 509–516.

Pool,W.G. *et al.* (1997a) Automated extraction of pure mass spectra from gas chromatographic/mass spectrometric data. *J. Mass Spectrom.*, **32**, 438–443.

Pool,W.G. *et al.* (1997b) Automated processing of GC/MS data: quantification of the signals of individual components. *J. Mass Spectrom.*, **32**, 1253–1257.

Saeed,A.I. *et al.* (2003) TM4: A free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

Saeed,A.I. *et al.* (2006) TM4 microarray software suite. *Methods Enzymol.*, **411**, 134–193.

Schauer,N. *et al.* (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett.*, **579**, 1332–1337.

Sinha,A.E. *et al.* (2004) Algorithm for locating analytes of interest based on mass spectral similarity in GC x GC-TOF-MS data: analysis of metabolites in human infant urine. *J. Chromatogr. A*, **1058**, 209–215.

Smith,C.A. *et al.* (2006) XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Spasić,I. *et al.* (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *BMC Bioinformatics*, **7**, 281.

Stein,S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.*, **10**, 770–781.

Stephanopoulos,G. *et al.* (2004) Exploiting biological complexity for strain improvement through systems biology. *Nat. Biotechnol.*, **22**, 1261–1267.

Sumner,L.W. *et al.* (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, **62**, 817–836.

Sumner,L.W. *et al.* (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**, 211–221.

Tikunov,Y. *et al.* (2005) A novel approach for non-targeted data analysis for metabolomics: large-scale profiling of tomato fruit volatiles. *Plant. Physiol.*, **139**, 1125–1137.

Trethewey,R.N. *et al.* (1999) Metabolic profiling: a Rosetta stone for genomics? *Curr. Opin. Plant Biol.*, **2**, 83–85.

Van den Dool,H. and Kratz,P.D. (1963) A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography. *J. Chromatogr.*, **11**, 463–471.

Vorst,O. *et al.* (2005) A non-directed approach to the differential analysis of multiple LCMS derived metabolic profiles. *Metabolomics*, **1**, 169–180.

Vreuls,R.J.J. *et al.* (1999) Gas chromatography-time-of-flight mass spectrometry for sensitive determination of organic microcontaminants. *J. Mircocolumn. Sep.*, **11**, 663–675.

Wagner,C. *et al.* (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry*, **62**, 887–900.