Published online 28 April 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W481–W484***
doi:10.1093/nar/gkn194

MassTRIX: mass translator into pathways

Karsten Suhre[1,2,*] and Philippe Schmitt-Kopplin[3]

[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, 85764 Neuherberg, [2]Department of Biology, University of Munich (LMU), Großhaderner Straße 2, 82152 Planegg-Martinsried and [3]Institute of Ecological Chemistry, Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

Received January 30, 2008; Revised March 26, 2008; Accepted April 2, 2008

ABSTRACT

Recent technical advances in mass spectrometry (MS) have brought the field of metabolomics to a point where large numbers of metabolites from numerous prokaryotic and eukaryotic organisms can now be easily and precisely detected. The challenge today lies in the correct annotation of these metabolites on the basis of their accurate measured masses. Assignment of bulk chemical formula is generally possible, but without consideration of the biological and genomic context, concrete metabolite annotations remain difficult and uncertain. MassTRIX responds to this challenge by providing a hypothesis-driven approach to high precision MS data annotation. It presents the identified chemical compounds in their genomic context as differentially colored objects on KEGG pathway maps. Information on gene transcription or differences in the gene complement (e.g. samples from different bacterial strains) can be easily added. The user can thus interpret the metabolic state of the organism in the context of its potential and, in the case of submitted transcriptomics data, real enzymatic capacities. The MassTRIX web server is freely accessible at http://masstrix.org

INTRODUCTION

Metabolomics is the systems level approach to the quest of understanding all relevant metabolic processes of an organism. The starting point of any metabolomics study is the measurement of the largest possible set of all naturally occurring organic compounds in a biological sample. It is assumed that the substances that are identified represent the principal metabolites of the organism under study. In contrast to genomics (yielding the proteins that an organism can make), transcriptomics (indicating the proteins an organism intends to make), and proteomics (giving the proteins that are made), metabolomics ideally represents a true endpoint to the classical chain of 'omics' technologies in the paradigm of life (DNA–RNA–protein–metabolites), by quantifying the outcome of the action of all regulatory and enzymatic processes that are active in a cell at any given time (1–4).

Recent technology advances in ionization techniques and mass spectrometry (MS), in particular high-resolution instruments based on Fourier transform (FT) technology (ion cyclotron resonance-FT/MS and Orbitrap) or the newest time of flight (TOF) MS systems, now allow high-precision measurements of single metabolite masses within an error range down to only a few parts per million (p.p.m.) (5–8). Sub-p.p.m. precision in ultrahigh resolution can even be reached in routine and full scan with higher field magnets in ICR-FT/MS (9). Accordingly, a number of algorithms have been developed that allow deriving the most likely bulk chemical formula for any given individual mass (10–13). It is well known that a mass precision as high as 1 p.p.m. may not always suffice for an unambiguous annotation of all mass peaks considering all possible elements (14), but may be sufficient with materials of known elementary composition (15) or in the context of restricted metabolite possibilities of a given organism, such as presented here.

In a metabolomics setting, it is essential to go beyond the simple derivation of bulk chemical formula, and to annotate the MS data in the context of the actual metabolic pathways of the organism under study, as some recent studies in plants and bacteria show (16–18). A limited number of software tools for this task are available in the form of downloadable program code (19), but to our knowledge no online web server presently provides the possibility to simply upload a high precision mass spectrum and to readily annotate all mass peaks as potential metabolites of a given organism from an actualized database, and then to map automatically the identified compounds onto the metabolic pathway maps of that organism, optionally including information on gene expression or gene coding potential. The MassTRIX web server that we present here responds to this task.

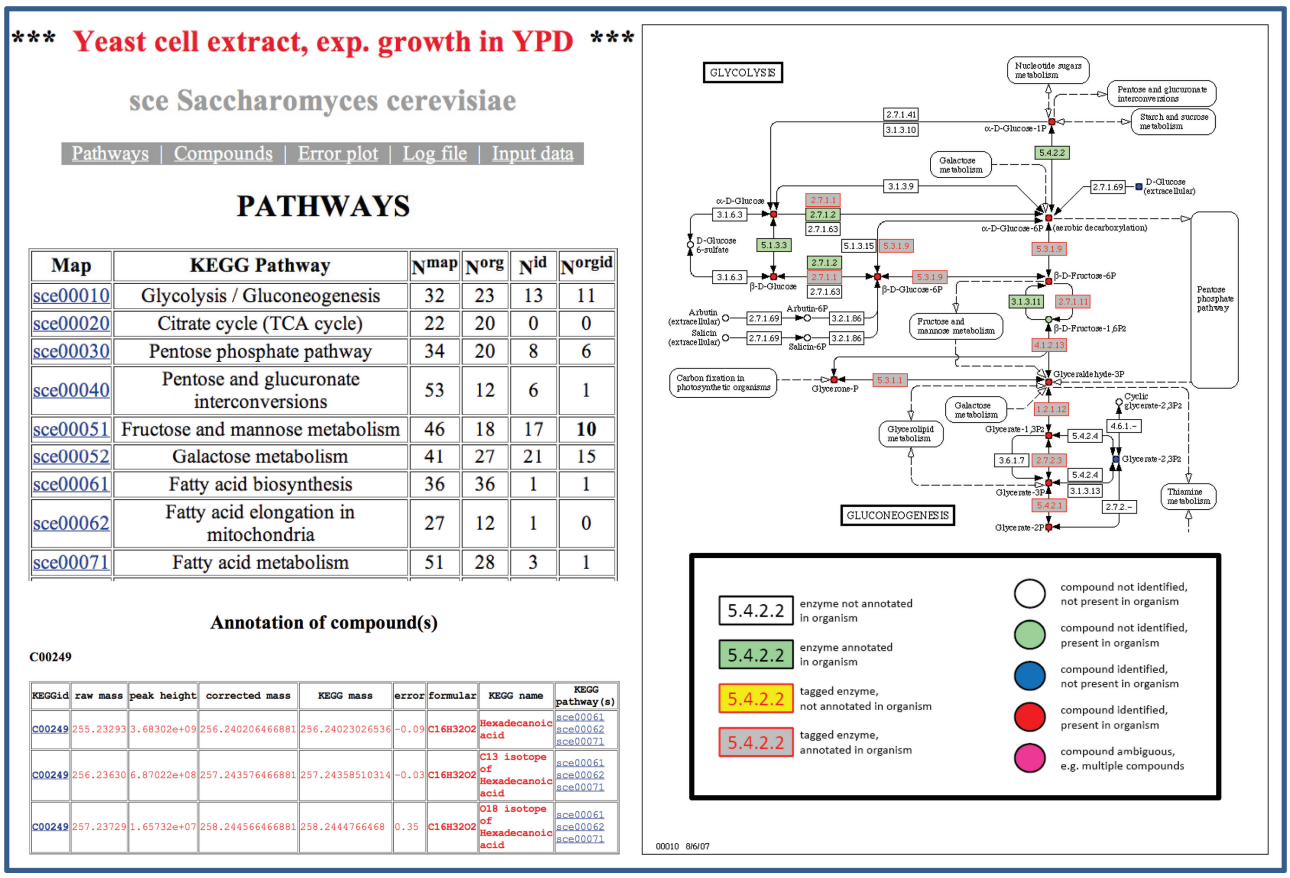
*To whom correspondence should be addressed. Tel: +49 89 3187 2627; Fax: +49 89 3187 3585; Email: karsten.suhre@helmholtz-muenchen.de

USING THE WEB SERVER

The required input data of MassTRIX is a mass peak list (i.e. mass–intensity pairs) from high-precision MS experiments, along with some parameters describing the dataset, such as the estimated error limit of the instrument and the ionization mode of the measurements. Typically, observations are from particular eukaryotic or prokaryotic cell extracts (e.g. using different gene knock-outs or growth conditions), but analysis of samples such as gut metabolomics have also been tested on our server. As basis for the annotation, an organism from the KEGG database (20) has to be selected. Optionally, a list of enzymes (EC-numbers) or genes (KEGG identifier) can be submitted to selectively highlight user-defined genes on the pathway maps.

MassTRIX then processes the submitted mass peak list by comparing the input experimental masses against all compounds of the KEGG chemical compound database, additionally including ^{13}C, ^{15}N and other isotopes, and optionally adding selected lipids with variable fatty acid chain lengths. Raw input masses from electrospray ionization (ESI) MS can be corrected on-the-fly for the addition or the abstraction of a proton (and optionally a Na^+ ion in positive mode). To cope with the requirement of very low measurement errors (in the sub-ppm range), exact masses of all KEGG compounds have been recomputed from the corresponding chemical formula using high-precision atomic mass data (21). MassTRIX then calls the KEGG/API (http://www.genome.jp/kegg/soap/) to generate pathway maps, where the identified compounds and genes are highlighted using different colors—thus differentiating between organism-specific and extra-organism items (Figure 1).

By default, all available KEGG pathway maps for the selected organism will be annotated (about 100–250, depending on the organism), but to speed up the computation, a subset of pathways can be specified by the user. A full run takes about 1 h; four parallel jobs can run at any time (this limit can be increased if the server usage warrants the upgrade, since most of the run time is spent in communication with the KEGG/API). The resulting colored pathway maps are fully clickable and cross-linked between different result pages from the server. These result pages include error plots, and pathway- and compound-specific pages, where also alternative annotations are visualized. Out-links to relevant KEGG enzyme



*** Yeast cell extract, exp. growth in YPD ***

sce Saccharomyces cerevisiae

Pathways | Compounds | Error plot | Log file | Input data

PATHWAYS

Map	KEGG Pathway	N^{map}	N^{org}	N^{id}	N^{orgid}
sce00010	Glycolysis / Gluconeogenesis	32	23	13	11
sce00020	Citrate cycle (TCA cycle)	22	20	0	0
sce00030	Pentose phosphate pathway	34	20	8	6
sce00040	Pentose and glucuronate interconversions	53	12	6	1
sce00051	Fructose and mannose metabolism	46	18	17	**10**
sce00052	Galactose metabolism	41	27	21	15
sce00061	Fatty acid biosynthesis	36	36	1	1
sce00062	Fatty acid elongation in mitochondria	27	12	1	0
sce00071	Fatty acid metabolism	51	28	3	1

Annotation of compound(s)

C00249

KEGGid	raw mass	peak height	corrected mass	KEGG mass	error	formular	KEGG name	KEGG pathway(s)
C00249	255.23293	3.68302e+09	256.240206466881	256.24023026536	-0.09	C16H32O2	Hexadecanoic acid	sce00061 sce00062 sce00071
C00249	256.23630	6.87022e+08	257.243576466881	257.24358510314	-0.03	C16H32O2	C13 isotope of Hexadecanoic acid	sce00061 sce00062 sce00071
C00249	257.23729	1.65732e+07	258.244566466881	258.2444766468	0.35	C16H32O2	O18 isotope of Hexadecanoic acid	sce00061 sce00062 sce00071

GLYCOLYSIS

GLUCONEOGENESIS

5.4.2.2 enzyme not annotated in organism

5.4.2.2 enzyme annotated in organism

5.4.2.2 tagged enzyme, not annotated in organism

5.4.2.2 tagged enzyme, annotated in organism

compound not identified, not present in organism

compound not identified, present in organism

compound identified, not present in organism

compound identified, present in organism

compound ambiguous, e.g. multiple compounds

00010 8/6/07

Figure 1. MassTRIX sample output from the yeast cell extract example job (http://masstrix.org/examples.html). The number of potentially identifiable and actually identified metabolites per pathway is reported, where N^{map} specifies the number of metabolites on the map, N^{org} gives the number of metabolites related to the organism, N^{id} is the number of experimentally identified metabolites on the map and N^{orgid} is the number of the identified metabolites that are related to the organism (top left). Clickable KEGG pathway maps (right) are colored according to the legend presented below. Metabolite nodes link to pages describing the mass peak annotations (bottom left), pathway nodes link to related annotated KEGG pathway maps on MassTRIX, while enzyme nodes link to the KEGG gene database. Auxiliary job information (input data, log files, error plot) can be accessed from all pages.

and compound pages are provided. A full documentation together with a list of frequently asked questions and a description of all options is maintained online at http://masstrix.org/doc.html.

INTERPRETING THE RESULTS

To illustrate the interpretation of a typical MassTRIX job, a mass spectrum from a cytosolic yeast cell extract under exponential growth, collected in positive scan mode on a Bruker-Daltonics APEXQ 12 Tesla ICR-FT mass spectrometer is provided as a work-through example at http://masstrix.org/examples.html. The input data for this test case are available through this web page, so that this test job can be rerun at any time. After completion of the job, two major starting points for interpretation are provided by MassTRIX (Figure 1): the '*Pathways*' page summarizes the number of identified metabolites on all available pathway maps, while the '*Compounds*' page gives a list of all metabolites that are annotated on any given pathway of the organism (here *Saccharomyces cerevisiae*). From these two major starting points, the user can navigate to the different maps and detailed annotations of the identified metabolites. Here it is important to note that in some cases multiple alternative annotations may be found, e.g. in cases where metabolites with identical bulk formula exist, such as fructose, mannose, allose and glucose 6-phosphates. In such a case, the pathway maps may be useful for deciding which of the annotated metabolites is likely to be present under the conditions of the study (here allose 6-phosphate is not a known metabolite of yeast in KEGG and may thus be ruled out). Other potential reasons to rule out some of the annotations could be related to too low concentrations of the metabolites to be seen with the used ionization and mass spectrometric technique, missing major isotopic mass peaks, or the isolated presence of some metabolites in the pathway maps. Multiple annotations may also occur when the masses of two different metabolites lie both within the error range of the observed mass peak. Here again, the pathway maps may be useful for making an educated guess about which metabolite is the most likely candidate.

In this concrete example, some specific biological facts can be readily gleaned from the MassTRIX output. For instance, as expected for yeast cells in exponential growth, a large number of metabolites in the glycolysis/gluconeogenesis pathway has been identified. These metabolites are preferentially connected to the highly expressed genes, indicated by red–gray enzyme boxes on the KEGG pathway map. On the other hand, the citrate cycle (TCA cycle) is shut off during yeast exponential growth. Accordingly, no metabolites have been identified that are linked to expressed genes of this pathway.

CONCLUDING REMARKS

The first purpose of the MassTRIX server presented here is the web-based analysis of MS-based metabolomics data, and its hypothesis-driven interpretation within the genomic context of the organism under study. To continually respond to its objectives, future developments of the MassTRIX web server shall include an automated preprocessing step of the input data, using for example heuristic filtering rules of molecular formulas (8,10,13), together with an inclusion of other metabolite databases, such as MetaCyc (22), the Human Metabolome Database (23) and the Database of Natural Products (24).

At this point, it is important to emphasize the limitations of this approach: we reported recently that high-resolution mass spectra reflect the isomer filtered complement of the entire space of molecular structures (9). An annotation such as the one proposed here thus associates experimental accurate mass (within an experimental error) with a limited number of bulk chemical formula (isomers), derived from the unique elementary composition space and restricted by the choice of the organism (and its annotated genome). The differentiation between isomers and the final metabolite identification can only be done on a case by case basis in further identification steps, using classical analytical chemistry approaches involving metabolite orthogonal separation, spectroscopy and further spectrometry together with chemical synthesis (25). An educated interpretation of the resulting pathway in the light of the genome of the organism thus remains the golden rule.

ACKNOWLEDGEMENTS

We thank Judith Glöckner-Pagel and Philipp Pagel for providing us with the yeast cell extract. Funding to pay the Open Access publication charges for this article was provided by the Helmholtz Zentrum München.

Conflict of interest statement. None declared.

REFERENCES

1. Nicholson,J.K., Connelly,J., Lindon,J.C. and Holmes,E. (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.*, **1**, 153–161.
2. Fiehn,O. (2002) Metabolomics—the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155–171.
3. Bino,R.J., Hall,R.D., Fiehn,O., Kopka,J., Saito,K., Draper,J., Nikolau,B.J., Mendes,P., Roessner-Tunali,U., Beale,M.H. *et al.* (2004) Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.*, **9**, 418–425.
4. Hirai,M.Y., Yano,M., Goodenowe,D.B., Kanaya,S., Kimura,T., Awazuhara,M., Arita,M., Fujiwara,T. and Saito,K. (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc. Natl Acad. Sci. USA*, **101**, 10205–10210.
5. Breitling,R., Pitt,A.R. and Barrett,M.P. (2006) Precision mapping of the metabolome. *Trends Biotechnol.*, **24**, 543–548.
6. Makarov,A., Denisov,E., Lange,O. and Horning,S. (2006) Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *J. Am. Soc. Mass Spectrom.*, **17**, 977–982.
7. Satoh,T., Sato,T. and Tamura,J. (2007) Development of a high-performance MALDI-TOF mass spectrometer utilizing a spiral ion trajectory. *J. Am. Soc. Mass Spectrom.*, **18**, 1318–1323.
8. Schmitt-Kopplin,P. and Hertkorn,N. (2007) Ultrahigh resolution mass spectrometry. *Anal. Bioanal. Chem.*, **389**, 1309–1310.
9. Hertkorn,N., Ruecker,C., Meringer,M., Gugisch,R., Frommberger,M., Perdue,E.M., Witt,M. and Schmitt-Kopplin,P. (2007) High-precision frequency measurements: indispensable tools

at the core of the molecular-level analysis of complex systems. *Anal. Bioanal. Chem.*, **389**, 1311–1327.
10. Kind,T. and Fiehn,O. (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinform.*, **8**, 105.
11. Ohta,D., Shibata,D. and Kanaya,S. (2007) Metabolic profiling using Fourier-transform ion-cyclotron-resonance mass spectrometry. *Anal. Bioanal. Chem.*, **389**, 1469–1475.
12. Katajamaa,M. and Oresic,M. (2007) Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A*, **1158**, 318–328.
13. Rosselló-Mora,R., Lucio,M., Peña,A., Brito-Echeverría,J., López-López,A., Valens-Vadell,M., Frommberger,M., Antón,J. and Schmitt-Kopplin,P. (2008) Metabolic evidences of biogeographic isolation of the extremophilic bacterium Salinibacter ruber. *The ISME Journal* (in press).
14. Kind,T. and Fiehn,O. (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinform.*, **7**, 234.
15. Hertkorn,N., Benner,R., Schmitt-Kopplin,P., Kaiser,K., Kettrup,A. and Hedges,I.J. (2006) Characterization of a major refractory component of marine organic matter. *Geochim. Cosmochim. Acta.*, **70**, 2990–3010.
16. Aharoni,A., Ric de Vos,C.H., Verhoeven,H.A., Maliepaard,C.A., Kruppa,G., Bino,R. and Goodenowe,D.B. (2002) Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *Omics*, **6**, 217–234.
17. Hirai,M.Y., Klein,M., Fujikawa,Y., Yano,M., Goodenowe,D.B., Yamazaki,Y., Kanaya,S., Nakamura,Y., Kitayama,M., Suzuki,H. *et al.* (2005) Elucidation of gene-to-gene and metabolite-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics. *J. Biol. Chem.*, **280**, 25590–25595.
18. Tang,Y., Pingitore,F., Mukhopadhyay,A., Phan,R., Hazen,T.C. and Keasling,J.D. (2007) Pathway confirmation and flux analysis of central metabolic pathways in Desulfovibrio vulgaris hildenborough using gas chromatography-mass spectrometry and Fourier transform-ion cyclotron resonance mass spectrometry. *J. Bacteriol.*, **189**, 940–949.
19. Jourdan,F., Breitling,R., Barrett,M.P. and Gilbert,D. (2008) MetaNetter: inference and visualization of high-resolution metabolomic networks. *Bioinformatics.*, **24**, 143–145.
20. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
21. Wapstra,A.H., Audi,G. and Thibault,C. (2003) The AME2003 atomic mass evaluation (I). Evaluation of input data, adjustment procedures. *Nuclear Phys. A.*, **729**, 129–336.
22. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **36**, D623–D631.
23. Wishart,D.S., Tzur,D., Knox,C., Eisner,R., Guo,A.C., Young,N., Cheng,D., Jewell,K., Arndt,D., Sawhney,S. *et al.* (2007) HMDB: the human metabolome database. *Nucleic Acids Res.*, **35**, D521–D526.
24. Buckingham,J. (ed.). (1993) C.H.C. *Dictionary of Natural Products,* 6th edn. Chapman & Hall, CRC Press, England.
25. Chen,J., Zhao,X., Fritsche,J., Yin,P., Schmitt-Kopplin,P., Wang,W., Lu,X., Häring,H.U., Schleicher,E.D., Lehmann,R. *et al.* (2008) Practical approach for the identification and isomer elucidation of biomarkers detected in a metabonomic study for the discovery of individuals at risk for diabetes by integrating the chromatographic and mass spectrometric information. *Anal Chem.*, **80**, 1280–1289.