BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 8 2005, pages 1635–1638
doi:10.1093/bioinformatics/bti236

Systems biology

GMD@CSB.DB: the Golm Metabolome Database

Joachim Kopka[1], Nicolas Schauer[1], Stephan Krueger[1], Claudia Birkemeyer[1], Björn Usadel[1], Eveline Bergmüller[2], Peter Dörmann[1], Wolfram Weckwerth[1], Yves Gibon[1], Mark Stitt[1], Lothar Willmitzer[1], Alisdair R. Fernie[1] and Dirk Steinhauser[1,*]

[1]Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany and [2]Institute of Plant Sciences, Swiss Federal Institute of Technology, 8092, Zurich, Switzerland

Received on October 20, 2004; revised on November 16, 2004; accepted on December 15, 2004
Advance Access publication December 21, 2004

ABSTRACT

Summary: Metabolomics, in particular gas chromatography–mass spectrometry (GC–MS) based metabolite profiling of biological extracts, is rapidly becoming one of the cornerstones of functional genomics and systems biology. Metabolite profiling has profound applications in discovering the mode of action of drugs or herbicides, and in unravelling the effect of altered gene expression on metabolism and organism performance in biotechnological applications. As such the technology needs to be available to many laboratories. For this, an open exchange of information is required, like that already achieved for transcript and protein data. One of the key-steps in metabolite profiling is the unambiguous identification of metabolites in highly complex metabolite preparations from biological samples. Collections of mass spectra, which comprise frequently observed metabolites of either known or unknown exact chemical structure, represent the most effective means to pool the identification efforts currently performed in many laboratories around the world. Here we present GMD, The Golm Metabolome Database, an open access metabolome database, which should enable these processes. GMD provides public access to custom mass spectral libraries, metabolite profiling experiments as well as additional information and tools, e.g. with regard to methods, spectral information or compounds. The main goal will be the representation of an exchange platform for experimental research activities and bioinformatics to develop and improve metabolomics by multidisciplinary cooperation.

Availability: http://csbdb.mpimp-golm.mpg.de/gmd.html

Contact: Steinhauser@mpimp-golm.mpg.de

Supplementary information: http://csbdb.mpimp-golm.mpg.de/

INTRODUCTION

The sequencing and annotation of whole genomes of various organisms (Goffeau *et al*., 1996; Blattner *et al*., 1997; The *Arabidopsis* Genome Initiative, 2000; Lander *et al*., 2001) facilitate the development of technology platforms to monitor the cellular inventory (Fiehn *et al*., 2000; Lockhart and Winzeler, 2000; Corbin *et al*., 2003). Since the dawn of genomic technology in the past decade and in conjunction with enhancing genomic information a vast amount of diverse data has been generated and released to the public community. The improving knowledge of gene functions in concurrence with global expression analyses allows phenotypes to be linked to their co-responding genomic data. However, our knowledge of the molecular basis of biological functions and their respective contribution to observed phenotypes is, as yet, relatively rudimentary. The recent mining and exploitation of data by multiparallel 'omics' technologies open up the possibility to gain comprehensive insight into the understanding of biological systems (Kitano, 2002; Oltvai and Barabási, 2002; Fernie *et al*., 2004). The flood of information obtained worldwide by scientists for this purpose urgently requires user-friendly public data access. In the past decades much progress has been made on the storage of information derived from the various levels of the cellular hierarchy. For instance, databases like BRENDA (Schomburg *et al*., 2004), KEGG (Kanehisa *et al*., 2004) or MetaCyc (Krieger *et al*., 2004) harbour information concerning metabolic pathways, chemical reactions including inventories of the genes and enzymes involved. Genomic databases, such as MIPS (Mewes *et al*., 2004), TAIR (Rhee *et al*., 2003) and TIGR (Quackenbush *et al*., 2000), provide public access to protein sequences based on whole genome analyses, maps of protein–protein interactions, protein localization and many further features. Developments in transcript profiling technologies have led to the adoption of uniform experimental platforms that are used worldwide. The widely shared experimental approach facilitated the establishment of expression profile related databases, such as the Stanford Microarray Database [SMD (Gollub *et al*., 2003)], TAIR or NCBI–GEO (Edgar *et al*., 2002). Similarly the availability of proteome data has led to the establishment of various databases [e.g. Swiss-Prot (Boeckmann *et al*., 2003)] or initiatives [e.g. HPI, (Hermjakob *et al*., 2004)] which focus on the functional annotation of proteins and standardization of protein data.

In contrast to the multitude of well established databases which comprise information on the genome, transcriptome and proteome, no attempt has been made to store the flood of data arising from metabolome analyses of biological samples. As already outlined, metabolites have an enormous diversity of chemical structures. These are identified and quantified using a wide range of technology platforms (Kopka *et al*., 2004). Thus there is an urgent need for publicly accessible metabolome databases that harbour underlying information. Here we describe the Golm Metabolome Database (GMD), an

*To whom correspondence should be addressed.

© The Author 2004. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

open access database for exchange and presentation of metabolomic and related information. In the current build GMD focuses on gas chromatography–mass spectrometry (GC–MS) (Roessner *et al*., 2000), the most advanced and widespread technology platform for metabolomics. The collected information (1) covers knowledge concerning analytical technologies and (2) harbours information that supports unequivocal metabolite identification. In addition, GMD provides access to stored metabolite profiles.

SYSTEMS OVERVIEW

Affiliation and implementation

The GMD platform is affiliated to CSB.DB, a comprehensive systems-biology database, which is hosted at the Max-Planck-Institute of Plant Molecular Physiology, Potsdam–Golm, Germany (Steinhauser *et al*., 2004). GMD complements the currently available transcriptional co-response databases and uses a similar system for data storage and handling.

Information on analytical technologies

The highly complex nature and the enormous chemical diversity of compounds obtained when analyzing the metabolome of organisms constitutes one of the main challenges in metabolomics (Oksman-Caldentey *et al*., 2004; Fernie *et al*., 2004). Current estimations vary. However, 4000–25 000 compounds may represent the metabolome of any given organism (Trethewey, 2004; Fernie *et al*., 2004). The plant kingdom is believed to comprise in excess of 200 000 metabolites (Fiehn, 2002; Trethewey, 2004). Highly diverse chemical characteristics in conjunction with the vast amount of potential compounds have profound implications for metabolite extraction and stability. Any given protocol for metabolome measurement thus represents a well tuned balance between accuracy and metabolite coverage. The GMD analytics pages allow access to expert knowledge on methods applied by the GMD contributors. Information on different technology platforms, publicly available methods, as well as contact information for individual knowledge exchange is included. Furthermore, an overview of the available resources is given for those scientists who intend to enter the field of experimental physiology and plan to set up a metabolomics facility.

Mass spectra and retention time index (MSRI) libraries

Following analytical measurements, data processing algorithms are applied to detect metabolic components in spectral data. The identification and characterization of the hundreds to thousands of metabolites obtained from diverse biological samples represents a major challenge in metabolomics. These identification efforts require large-scale processing of pure standard substances to generate customized spectral libraries that can be used for subsequent identification of hitherto unknown metabolic components from spectral data. To overcome the current limitation of customized mass spectral libraries that need to be maintained by each laboratory the GMD mass spectra information pages are developed to exchange information. In detail, we started to disseminate the underlying evidences that support metabolite identification in complex GC–MS profiles from diverse biological sources. The MSRI web platform provides access to customized MSRI libraries, which were generated using identical capillary GC columns and settings using two different electron impact ionization GC–MS technologies, namely quadrupole GC–MS (Fiehn *et al*., 2000; Roessner *et al*., 2000) and GC–TOF (time-of-flight)–MS (Wagner *et al*., 2003; Weckwerth *et al*., 2004). Currently, five downloadable libraries are available, which may be imported into the NIST02 mass spectral search program or AMDIS, a technology platform independent automated mass spectral deconvolution and identification system (National Institute of Standards and Technology, Gaithersburg, MD, USA). The above libraries are split according to the technology platform and the degree of manual mass spectral curation. The Q_MSRI and T_MSRI libraries contain mass spectral tags (MSTs), which were either generated on four identically configured quadrupole GC–MS systems (Q_MSRI) or on a single time-of-flight system (T_MSRI). Mass spectral libraries, which exclusively consist of manually evaluated, identified or classified MSTs are assigned to ID-libraries. In contrast, libraries which were generated by automated deconvolution were assigned to NS libraries, indicative of the non-curated mode of construction. The currently available libraries cover data from mammals, yeast, corynebacterium, model plants, such as crop plants and related wild species, as well as required non-sample controls. These libraries currently feature more than 2000 evaluated mass spectra from the two technology platforms which represent 1089 non-redundant and 360 identified MSTs.

The metabolite profiling platform (GMD profiles)

The vast amount of complex data obtained from metabolite profiling experiments in conjunction with the ongoing developments on analytical technologies require the public availability of these data for cross-comparison and cross-experiment analysis. According to these demands we started to present metabolic fingerprinting and metabolite profiling experiments, which can be currently searched by compound names or browsed by a list of experiments.

For the exchange of the highly complex experimental background information and data from metabolite profiling experiments we implemented the MIAMET description as suggested by Bino *et al*. (2004). For future implementations and development of the GMD platform recently made advances in database modelling and insight into the architecture of metabolomics data (ArMet) will prove to be highly important (Jenkins *et al*., 2004).

IMPLEMENTATION AND QUERY OVERVIEW

Content browsing and queries

The GMD content can be explored by browsing the HTML content through lists or a simple site map, represented as a hierarchical tree, which is linked to the available second level of HTML pages. Information regarding downloadable MSRI libraries as well as related Supplementary information, such as technologies, method descriptions and acknowledgements, are made accessible. Both, the MSRI libraries as well as the currently integrated metabolite profiling experiments are presented in table format, which provides links to associated detailed information.

A more sophisticated way to explore the GMD content is offered through the available query pages. Currently, five different types of queries are implemented which can all be accessed by the GMD site map.

MSRI compound search

The compound search tool allows searching by compound name and provides access to the linked mass spectral information harboured at GMD. Various filter options can be applied to restrict the query results, e.g. to the available technology platforms, particular libraries or methods. The retrieved mass spectral entries are presented as a table which contains basic mass spectral information for a particular compound, such as compound role, i.e. metabolite or internal standard, observed retention time index (RI) and technology platform. This basic information can be sorted upon user invocation. All information is linked to the detailed physicochemical characteristics of the available mass spectra, which are represented as a mass spectrum chart. This final level of information facilitates the identification of compounds in profile analyses. The in-depth mass spectral information encompasses in addition (1) the recommended quantifier and qualifier masses, (2) access to available replicate mass spectra of the same compound and (3) a direct link to the mass spectrum search and comparison tool.

MSRI mass spectrum search

For analysis of mass spectra that are present in the libraries but can also be user-submitted we implemented a query tool which allows comparison with all available curated mass spectra of our libraries. Mass spectra may be submitted in either NIST02 or AMDIS format (Ausloos *et al*., 1999; Stein, 1999). The search is performed by computing the fragment-intensity agreement, measured as dynamically normalized Euclidean distance [Euclid], as S12 [s12] index (Gower and Legendre, 1986), Hamming (1950) and Jaccard (1908) distance. The result set is presented as a sortable HTML table containing information such as the rank, the identifier for each spectrum, the RI, the method information, the compound name in case of identified metabolites and all computed similarity measures. All types of information can be used for sorting. Moreover, additional criteria for comparison are given based on absolute RI differences to (1) the observed RI as provided by an optional user input and (2) as calculated relative to the best hit. If available, occurrence of qualifier as well as quantifier masses is considered. A head-to-tail plot of the query and selected hit spectra can be invoked. Depending on the chosen sorting a colour-coded graphical representation of the ten best hits is generated below the result table. The graphical output is similar to a typical BLAST (Altschul *et al*., 1990) result. The ratio plot mirrors the occurrence of the masses and their co-responding ratios of intensities in comparison to the query spectrum. The result table can be downloaded by an exporter function as a tab-delimited and zip-coded file. The file contains all data presented in HTML table and in addition all returned mass spectra of the query. Various filter options, especially restriction to a predefined RI window or set of major fragments, can be invoked by the user to limit the search to relevant results. The set of implemented tools is complementary to those available within the NIST02 software.

MSRI customized library generation

In extension to the precompiled MSRI download libraries GMD allows the generation of user-customized mass spectral subsets from the full repository of curated mass spectral entries. These subsets can be downloaded as a zip-coded text file and treated like our precompiled MSRI libraries (see above). The search input is currently restricted to MPIMP-Ids, which can be obtained through the abovementioned queries or by using the compound name converter (see below). We suggest limiting of results according to the GC–MS technology platform or the analytical methods used in order to obtain the curated spectra.

Profile compound search

As mentioned above GMD has started to integrate a first set of metabolite profiling experiments which were generated with a quadrupole GC–MS technology platform. Currently, 69 profiles of nine replicate sets are included describing metabolic changes under different light conditions. The profiles can be queried by compound name and allows searches for the changes in compound levels. Various filter options are available to restrict computation to high-quality mass traces by using the default or user modified values. Moreover, the user can select between parametric or non-parametric statistics for the dynamic computation of the treatment versus control comparisons. The result set covers information on experimental background, performed comparisons as well as information on significance of the observed differences. Furthermore, treatment versus control ratios are given and colour coded to mark decreases or increases. In analogy to the Affymetrix oligonucleotide technology platform we use different masses as representatives for any particular compound. All used masses are represented in the result tables and are checked for co-responding behaviour across the full experimental dataset. Future updates will connect metabolite profiles of GMD to the visualization software tools MapMan (Thimm *et al*., 2004) and PaVESy (Ludemann *et al*., 2004).

Compound name converter

Because of the different usage of compound names and identifiers in the publicly available databases we implemented a converter which allows converting of user compound names to MapMan names for a stand-alone visualization of the results with the MapMan software (Thimm *et al*., 2004) or to MPIMP-Ids for customized library generation.

OUTLOOK

GMD will frequently be updated with new mass spectra, metabolite identifications, mass spectral libraries of biological samples and metabolite profiling experiments. GMD is intended as a repository for experiments performed at the Max-Planck-Institute of Molecular Plant Physiology and for data made available through collaborating scientists. We offer our already well-characterized GC–MS technology platforms specifically for cooperations on metabolite identification in complex biological samples. As suggested by Bino *et al.* (2004) we envision to share biological samples and metabolite identifications between laboratories engaged in GC–MS metabolite profiling. Thus we provide a public platform for future advances and developments in metabolomic science. In-depth analysis and understanding of metabolome data at systems level will require a multidisciplinary effort, especially integration of proteome and transcriptome data. Such interdisciplinary cooperation and data mining is in preparation and in the case of steady state transcript analysis already in place (Steinhauser *et al*., 2004). We are convinced that GMD will represent a crucial building block for CSB.DB (http://csbdb.mpimp-golm.mpg.de). CSB.DB, a comprehensive systems-biology database project, will harbour and allow joined access to metabolome, proteome and transcriptome data. Thus CSB.DB will develop into a highly useful and informative public

resource for researchers focusing on experimental biology as well as for computational biology and bioinformatics.

ACKNOWLEDGEMENTS

We appreciate the work of all scientists, who contributed samples and submitted mass spectral information or metabolite profiling experiments to GMD. Detailed acknowledgements and affiliations are made accessible through GMD (http://csbdb.mpimp-golm.mpg.de/gmd.html). We are grateful to the Max-Planck-Institute of Molecular Plant Physiology and the Max-Planck-Society for long-standing and continuous support of the Golm Metabolome Database.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignments search tool. *J. Mol. Biol.*, **215**, 403–410.

Ausloos,P., Clifton,C.L., Lias,S.G., Mikaya,A.I., Stein,S.E., Tchekhovskoi,D.V., Sparkman,O.D., Zaikin,V. and Zhu,D. (1999) The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.*, **10**, 287–299.

Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.

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. *Trend Plant Sci.*, **9**, 418–425.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Corbin,R.W., Paliy,O., Yang,F., Shabanowitz,J., Platt,M., Lyons,C.E.,Jr, Root,K., McAuliffe,J., Jordan,M.I., Kustu,S. *et al*. (2003) Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc. Natl Acad. Sci. USA*, **100**, 9232–9237.

Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Fiehn,O., Kopka,J., Dormann,P., Altmann,T., Trethewey,R.N. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, **18**, 1157–1161.

Fiehn,O. (2002) Metabolomics—the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155–171.

Fernie,A.R., Trethewey,R.N., Krotzky,A.J. and Willmitzer,L. (2004) Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.*, **5**, 763–769.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al*. (1996) Life with 6000 Genes. *Science*, **274**, 546–567.

Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. *et al*. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94–96.

Gower,J.C. and Legendre,P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.*, **3**, 5–48.

Hamming,R.W. (1950) Error detecting and error correcting codes. *Bell Syst. Tech. J.*, **9**, 147–160.

Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al*. (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.

Jaccard,P. (1908) Nouvelles recherches sur la distribution florale. *Bull Soc. Vaud Sci. Nat.*, **44**, 223–270.

Jenkins,H., Hardy,N., Beckmann,M., Draper,J., Smith,A.R., Taylor,J., Fiehn,O., Goodacre,R., Bino,R., Hall,R. *et al*. (2004) A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.*, **22**, 1601–1606.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.

Kopka,J., Fernie,A., Weckwerth,W., Gibon,Y. and Stitt,M. (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.*, **5**, 109.

Krieger,C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **32**, D438–D442.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., Fitzttugh,W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Lüdemann,D., Weicht,D., Selbig,J. and Kopka,J. (2004) PaVESy: pathway visualization and editing system. *Bioinformatics*, **20**, 2841–2844.

Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al*. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.

Oksman-Caldentey,K.-M., Inzé,D. and Orešič,M. (2004) Connecting genes to metabolites by a systems biology approach. *Proc. Natl Acad. Sci. USA*, **101**, 9949–9950.

Oltvai,Z.N. and Barabási,A.-L. (2002) Life's Complexity Pyramid. *Science*, **298**, 763–764.

Quackenbush,J., Liang,F., Holt,I., Pertea,G. and Upton,J. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.*, **28**, 141–145.

Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al*. (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.

Roessner,U., Wagner,C., Kopka,J., Trethewey,R.N. and Willmitzer,L. (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. *Plant J.*, **23**, 131–142.

Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.

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.

Steinhauser,D., Usadel,B., Luedemann,A., Thimm,O. and Kopka,J. (2004) CSB.DB: a comprehensive systems–biology database. *Bioinformatics*, **20**, 3647–3651.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Thimm,O., Blasing,O., Gibon,Y., Nagel,A., Meyer,S., Kruger,P., Selbig,J., Muller,L.A., Rhee,S.V. and Stitt,M. (2004) MAPMAN: a user-driven tool to display genomics datasets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914–939.

Trethewey,R.N. (2004) Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opin. Plant Biol.*, **7**, 196–201.

Wagner,C., Sefkow,M. and Kopka,J. (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.

Weckwerth,W., Loureiro,M.E., Wenzel,K. and Fiehn,O. (2004) Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl Acad. Sci. USA*, **101**, 7809–7814.