Plant Biotechnology 26, 479–483 (2009)

Bioinformatics Note

Visualization of metabolite identifier information

Fumio Matsuda[1], Henning Redestig[1], Yuji Sawada[1], Yoko Shinbo[2], Masami Yokota Hirai[1,3], Shigehiko Kanaya[1,2], Kazuki Saito[1,4,*]

[1] RIKEN Plant Science Center, Yokohama, Kanagawa 230-0045, Japan; [2] Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0101, Japan; [3] Japan Science and Technology Agency, CREST, Kawaguchi, Saitama 332-0012, Japan; [4] Graduate School of Pharmaceutical Sciences, Chiba University, Chiba 263-8522, Japan

* E-mail: ksaito@psc.riken.jp Tel: +81-45-503-9442 Fax: +81-45-503-9489

Received September 8, 2009; accepted October 5, 2009 (Edited by H. Suzuki)

Abstract In this study, we used a graphical representation to integrate, visualize, search, and analyze information on metabolite identifiers. By considering the links between metabolite identifiers described in the metabolite databases to be the edges between vertices in a graph, several metabolite databases can be integrated into a database without defining a new metabolite identifier code. The graphical visualization of metabolite identifier network enables us to understand the meaning of each metabolite identifier and their relationship with associated identifiers. The projection of actual metabolome data on the pathway map was also attained by using the converter function of the metabolite identifier database. We demonstrated that other metabolite-related information, such as chemical ontology and species-metabolite relationship, can be incorporated into the network, and performed an analysis of plant species-alkaloid ontology relationship.

Key words: Graph theory, metabolite identifier, plant metabolomics.

Special Issue

Metabolites in plants are addressed by using multiple identifiers since there is no definite system for metabolite notation, which hampers the practical use of metabolite information in the research of plant metabolism. In addition to many synonyms (names) of an identical metabolite, accession codes or registration numbers of several metabolite databases have been used, including the chemical abstract service (CAS) of American Chemical Society, Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) (Kanehisa et al. 2008), chemical entities of biological interest (ChEBI, http://www.ebi.ac.uk/chebi/) (Degtyarenko et al. 2008), and a comprehensive species-metabolite relationship database (KNApSAcK, http://kanaya.naist.jp/KNApSAcK_Family/) (Shinbo et al. 2006; Takahashi et al. 2008). Since the datasets are managed by using their own identifiers, the relationship among those metabolite identifiers should be understood for an exact notation of plant metabolites. However, the relationship is not simple, because the name of metabolite is often ambiguous, and incompatible classification rules are employed among the databases. For example, although "glutamate" does not specify the stereochemistry of metabolites, it is often used to address L-isomers. Furthermore, L-glutamate and L-glutamic acid have distinct identifiers in ChEBI (CHEBI:14321 and CHEBI:18237), whereas KEGG compound considered them to be an identical metabolite (KEGG:C00025). In addition, there are various types of metabolite-related information, such as literature-reported species-metabolite relationship data provided by KNApSAcK, metabolic pathway information by KEGG, and a chemical ontology system constructed by ChEBI, which suggest that a technical improvement is required to process the metabolite identifier information. Management of metabolite identifiers is a key technology for performing many metabolomics data analysis tasks such as dataset integration and contextual interpretation. We recently developed a software solution called MetMask to automatically integrate, map, and convert different metabolite identifiers (Redestig et al. submitted, http://sourceforge.net/projects/metmask/). MetMask can construct consensus metabolite mappings from a wide range of sources such as in-house reference libraries and external databases. For data analysis, it is essential that metabolite identifier mappings are *well-formed* so that conversions are commutative and constrained to be unique with each identifier mapping to exactly one metabolite. To fulfill these conditions, MetMask reorganizes the original data, which causes a loss of information. In order to compensate for this shortcoming, we wished to illustrate and explore the information stored in the original databases available and therefore developed a straightforward graph-based

This article can be found at http://www.jspcmb.jp/

approach to visualize and integrate metabolite identifier connections in this study.

The graph representation of a complex network has been employed to visualize and analyze biological data such as protein-protein interaction (Altaf-Ul-Amin et al. 2006) and gene coexpression (Saito et al. 2008). It should be noted that the relationship among metabolite identifiers also shapes networks, because an entry of metabolite databases can be regarded as a list of links between metabolite identifiers. For example, an entry of L-tyrptophan in KEGG compound (KEGG:C00078) is expressed as follows (Figure 1a).

KEGG:C00078 (links to) L-tryptophan
KEGG:C00078 (links to) Tryptophan
KEGG:C00078 (links to) $C_{11}H_{12}N_2O_2$
KEGG:C00078 (links to) CAS:73-22-3
KEGG:C00078 (links to) PubChemSID:3378
KEGG:C00078 (links to) ChEBI:16828

The data structure corresponds to a small, undirected graph by considering the link between two objects as an edge between two vertices (Figure 1b). We merged the small graphs derived from multiple databases and produced a complex network describing the relationship among the metabolite identifiers (Figure 1c). In order to investigate the performance of the graph-based approach, we used the MySQL 5.1 software (Sun Microsystems, USA) and created a database incorporating the link information from KEGG compound (2009/8/6 version), KNApSAcK (version 1.200), ChEBI (Release 59.0) data, and a standard compound list maintained by Platform of RIKEN Metabolomics (PRIMe; Akiyama et al. 2008). We distinguished the metabolite identifiers from different databases by adding prefixes such as CAS, KEGG, ChEBI, KNApSAcK, and PRIMe. The constructed dataset contains 232,264 vertices and 313,832 edges. We performed all analyses using in-house Perl scripts, and visualized the metabolite identifier networks with the Cytoscape software (version 2.6.0; Shannon et al. 2003).

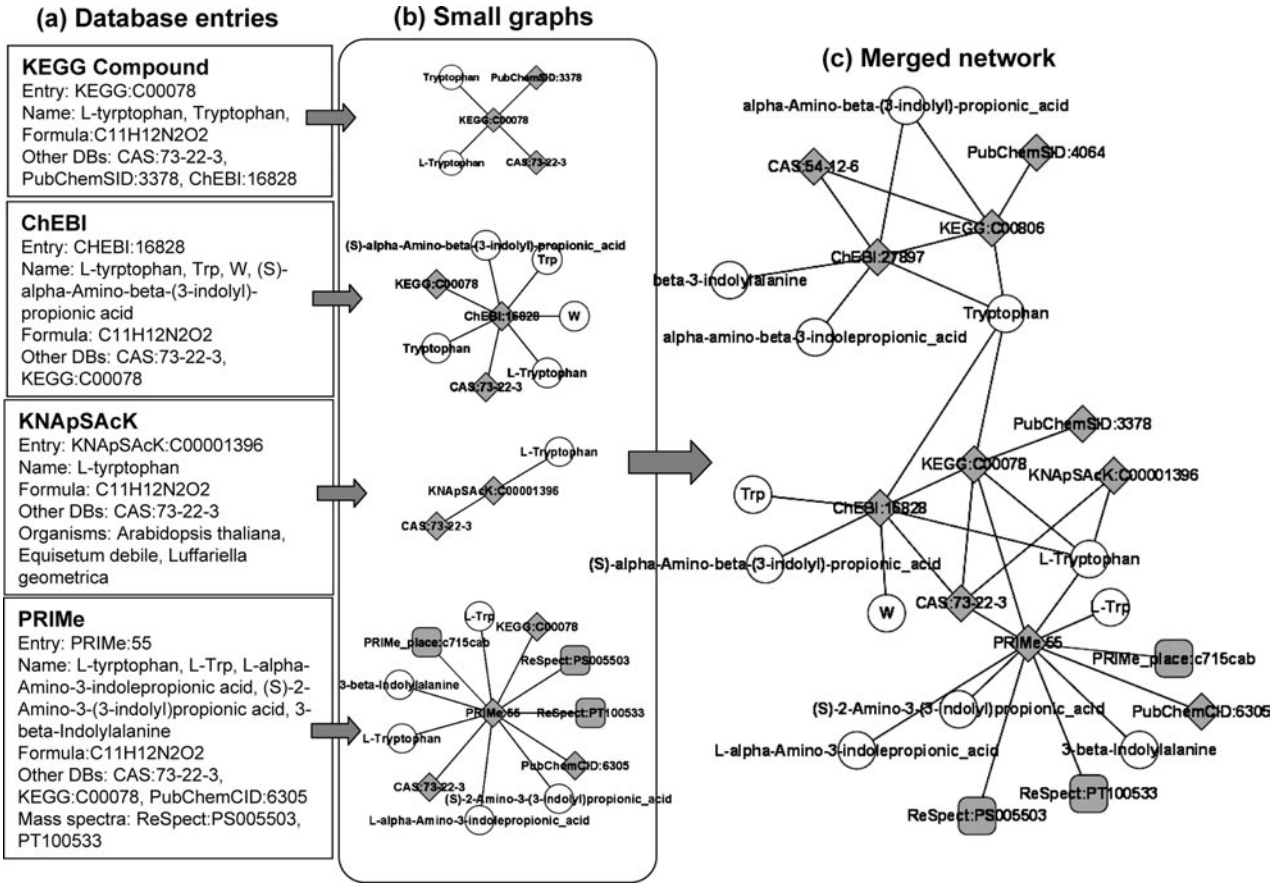

Figure 1. Graphical representation of the relationship among metabolite identifiers. (a) Entries of L-tryptophan in metabolite database, KEGG compound (Kyoto Encyclopedia of Genes and Genomes; http://www.genome.jp/kegg/), chemical entities of biological interest (ChEBI, http://www.ebi.ac.uk/chebi/), a comprehensive species-metabolite relationship database (KNApSAcK; http://kanaya.naist.jp/KNApSAcK_Family/), and a standard compound list maintained by Platform of RIKEN Metabolomics (PRIMe). (b) Small graph representing the link information determined by each database. Vertices of *open circle*, *filled diamond*, and *filled box* represent metabolite names, identifier codes, and additional information, respectively. Edges indicate the presence of a link between two vertices. (c) Merged network of metabolite identifiers connected to "L-tryptophan" obtained by using the breadth-first search algorithm. The information of racemic form of tryptophan was also obtained by the procedure.

Graphical visualization of the relationship among metabolite identifiers

Figure 1c shows a closed graph of metabolite identifiers connected to "L-tryptophan," obtained by the breadth-first search method (Golovin and Henrick 2009). The graphical representation simplifies the understanding of the relationship among identifiers of tryptophan since the meaning of each identifier could be elucidated from the context of neighbor vertices. For example, the two KEGG codes KEGG:C00078 and KEGG:C00806 obviously correspond to L-tryptophan and the racemic form of tryptophan, respectively. The function enables us to find a set of suitable identifiers addressing a metabolite of interest without checking multiple databases.

Identifier associations

Because the breadth-first search is complete, the graph

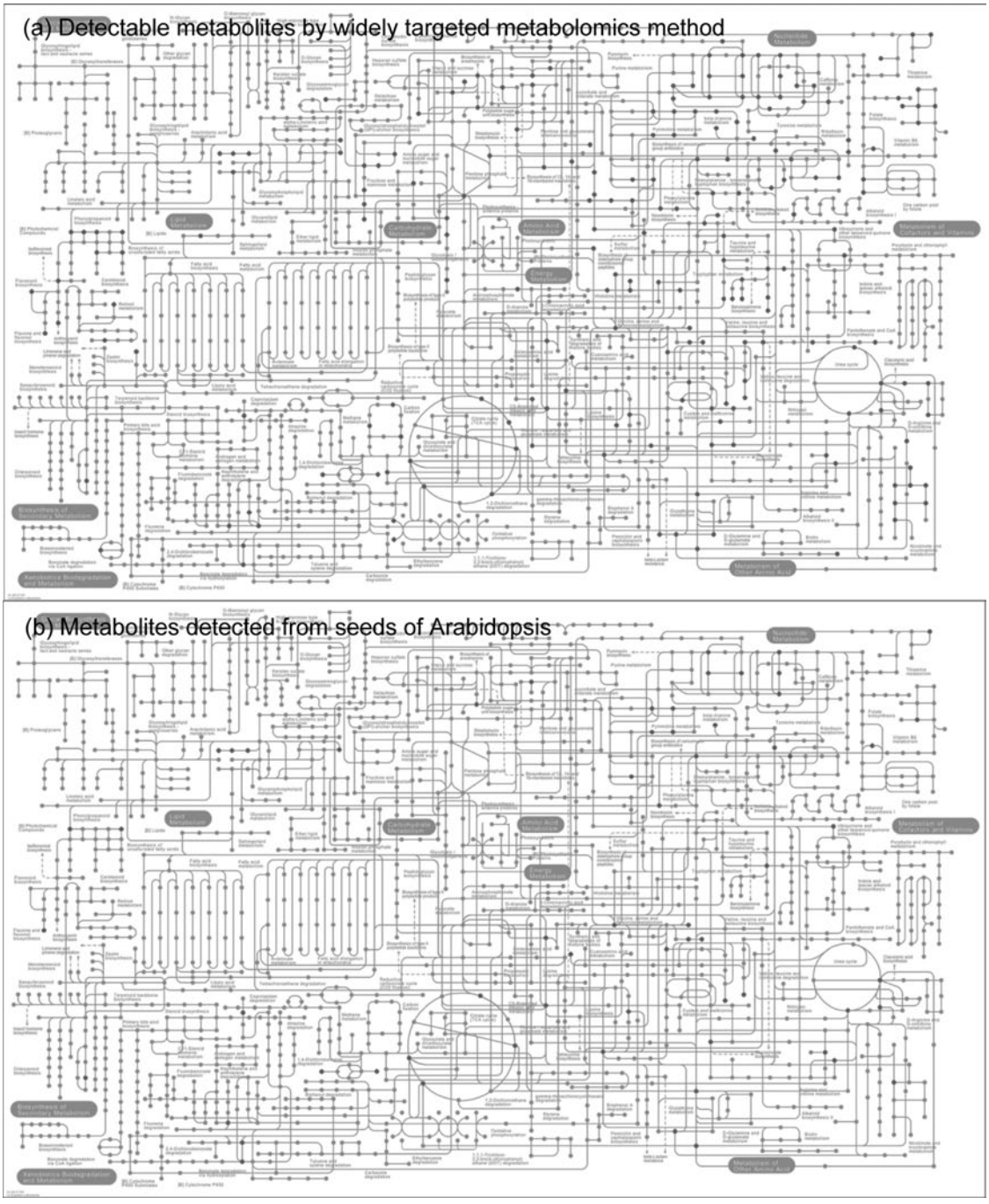

Figure 2. Visualization of metabolome data on KEGG pathways map. The names of metabolites were converted to the corresponding KEGG code by using the converter function of the developed database and then projected on the KEGG pathway map (http://www.genome.jp/kegg/pathway.html). (a) All metabolites detected by using the widely targeted metabolomics method (Sawada et al. 2009). (b) Metabolites detected from seeds of Arabidopsis.

shown in Figure 1c includes all identifiers connected to "L-tryptophan," which means that any corresponding identifier to "L-tryptophan" could be identified from the graph. In other words, the database can act as a converter across metabolite identifiers. If more than two candidate identifiers are found (e.g., KEGG:C00078 and KEGG:C00806 are candidates for KEGG code of a query "L-tryptophan"), the closer identifier from the query in the graph (KEGG:C00078) would be the most plausible candidate (although it was not guaranteed). By using the converter function, the metabolite identifiers in the actual metabolome data obtained from the 50 seeds of Arabidopsis by using the widely targeted metabolomics method (Sawada et al. 2009) were successfully converted to the corresponding KEGG codes. The converted KEGG codes made it possible to project the matabolome data on the KEGG pathway map by using the KEGG database function (http://www.genome.jp/kegg/pathway.html) as shown in Figure 2. The result indicated that various types of metabolites were produced in the seeds of Arabidopsis. The visualization of metabolome data using the converter function and pathway maps is able to help an intuitive interpretation of the result of metabolome analysis.

The network shown in Figure 1c contains two KEGG identifiers in a closed graph, suggesting that the network consists of identifiers of two different metabolites. The conflict of the metabolite identifiers can be checked by using a Union-Find algorithm (Bader et al. 2001), which has been developed to examine the connectivity of two vertices in a graph. An analysis of the connectivity of all pairs of KEGG identifiers indicated that 4.6% of the KEGG identifiers were conflicted. One reason for this relatively high conflict rate is the existence of ambiguous identifiers such as "tryptophan" which bridges two clusters of metabolite identifiers as shown in Figure 1c.

Incorporation of meta-information

The graph-based strategy allows the incorporation of any

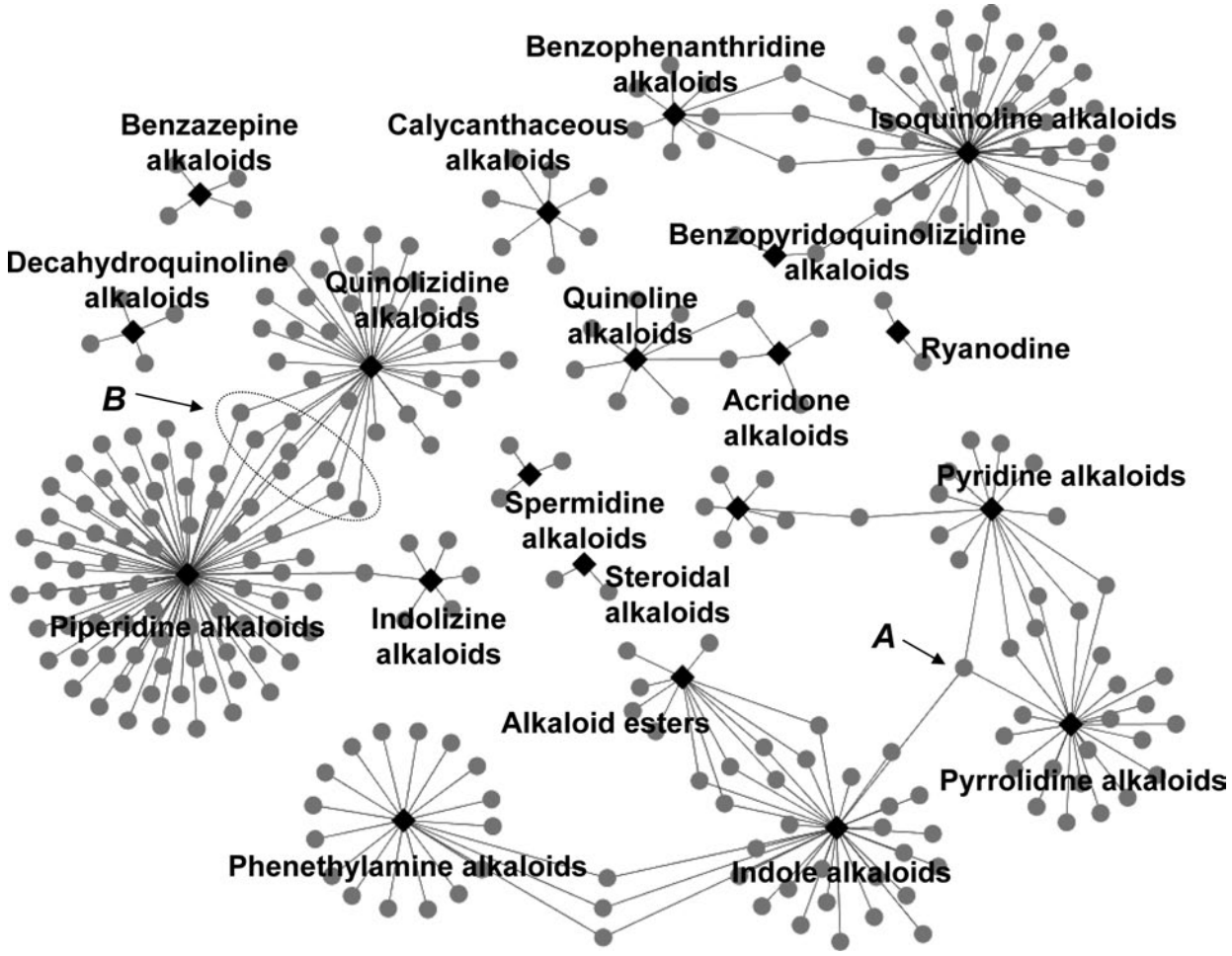

Figure 3. Networks representing the relationship between species-alkaloid ontologies. ChEBI codes of metabolites belonging to each chemical ontology terms of alkaloid defined by ChEBI (*filled diamonds*) were converted to the corresponding KEGG codes, and then the plant species (*filled circles*) producing those alkaloids were found. Edges indicate that the plant species produce alkaloids in the ontology. Symbol A denotes the plant species *Nicotiana tabacum*, which produces alkaloids of indole, pyridine, and pyrrolidine. Symbol B denotes the eight species producing alkaloids of both piperidine and quinolizidine.

type of metabolite-related information. For instance, the vertex assigned "PRIMe_place:c715cab" in Figure 1c indicates that a bottle of this reagent is stored in a cabinet at room C715. The vertices "ReSpect:PT100533" and "ReSpect:PS005503" represent accession codes of tandem mass spectral data of L-tryptophan in the RIKEN MSn spectral database for phytochemicals (ReSpect, http://spectra.psc.riken.jp/MassBank/). In addition, we incorporated the chemical ontology data provided by ChEBI and metabolite-species relationship from KNApSAcK into the database by introducing a directed graph to describe the implication relationship between metabolite identifiers and the metadata. ChEBI defines the chemical ontology terms of plant secondary metabolites such as flavonoid, glucosinolate, alkaloid, and those subgroups such as "isoquinoline alkaloids" and "piperidine alkaloids". By using the information, relationships between plant species-alkaloid ontology were investigated. We converted the ChEBI codes of metabolites belonging to each ontology term to the corresponding KEGG codes by the converter function described earlier, and then we found the plant species producing those alkaloids. Figure 3 shows a network of plant species-alkaloid ontology relationships. Although a majority of plant species belonged to only a single alkaloid type, there are some species producing two or more different types of alkaloids. For instance, *Nicotiana tabacum* (symbol A in Figure 3) produced tryptamine (a member of indole alkaloid) and nicotine (which belongs to both pyridine and pyrrolidine alkaloids). In addition, 8 Fabaceae species (symbol B in Figure 3; *Cytisus austriacus*, *Cytisus balansae*, *Cytisus eriocarpus*, *Genista lydia*, *Genista tinctoria*, *Lupinus pilosus*, *Spartidium saharae*, *Thermopsis chinensis*) are capable of producing two different types of alkaloids, (+)-ammodendrine (KNApSAcK:C00002014) and sparteine (KNApSAcK:C00002236), which are members of piperidine and quinolizidine alkaloids, respectively. The abovementioned eight species belong to different genus, indicating that more species in Fabaceae family are likely to produce the above two types of alkaloids. Although details of those biosyntheses remain unclear, our result suggests that piperidine and quinolizidine moiety in these alkaloids are biosynthesized from common precursor cadaberine, which is a decarboxylation product of lysine (Hartmann et al. 1980).

The graph-based approach is an automatic and information-lossless method to integrate metabolite identifier data without defining new metabolite identifier codes (Figure 1b). The graphical representation of metabolite identifier networks enables us to understand the meaning of metabolite identifiers (Figure 1c) and to perform metabolite identifier conversions (Figure 2). We also demonstrated that the graph-based method has the ability to integrate various metabolite-related information (Figures 1c, 3) and showed that techniques derived from the graph theory can be applied to analyze metabolite identifier networks. Furthermore, application of the graph-based method is expected to facilitate an advanced mining of metabolic data by integrating, visualizing, searching, and analyzing metabolite identifier information.

Acknowledgements

The authors wish to thank Koji Takano (RIKEN Plant Science Center) for technical support. This work was partly supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Genomics for Agricultural Innovation, NVR-0005) and JST/CREST (Project: "Elucidation of Amino Acid Metabolism in Plants based on Integrated Omics Analyses").

References

Akiyama K, Chikayama E, Yuasa H, Shimada Y, Tohge T et al. (2008) PRIMe: a Web site that assembles tools for metabolomics and transcriptomics. *In Silico Biol* 8: 339–345

Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S (2006) Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics* 7: 207

Bader DA, Moret BM, Yan M (2001) A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J Comput Biol* 8: 483–491

Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, et al. (2008) ChEBI: a database and ontology for chemical Entities of biological interest. *Nucl Acids Res* 36: D344–350

Golovin A, Henrick K (2009) Chemical substructure search in SQL. *J Chem Inform Modeling* 49: 22–27

Hartmann T, Schoofs G, Wink M (1980) A chloroplast-localized lysine decarboxylase of Lupinus polyphyllus: the first enzyme in the biosynthetic pathway of quinolizidine alkaloids. *FEBS lett* 115: 35–38

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. *Nucl Acids Res* 36: D480–484

Saito K, Hirai MY, Yonekura-Sakakibara K (2008) Decoding genes with coexpression networks and metabolomics—'majority report by precogs'. *Trends Plant Sci* 13: 36–43

Sawada Y, Akiyama K, Sakata A, Kuwahara A, Otsuki H, Sakurai T, Saito K, Hirai MY (2009) Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. *Plant Cell Physiol* 50: 37–47

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498–2504

Shinbo Y, Nakamura Y, Altaf-Ul-Amin M, Asahi H, Kurokawa K, et al. (2006) KNApSAcK: a comprehensive species-metabolite relationship database. In: Saito K, Dixon RA, Willmitzer L (eds) *Biotechnology in Agriculture and Forestry 57 Plant Metabolomics*. Springer, Berlin, pp 165–181

Takahashi H, Kai K, Shinbo Y, Tanaka K, Ohta D, et al. (2008) Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry. *Anal Bioanal Chem* 391: 2769–2782