KEGG and GenomeNet Resources for Predicting Protein Function from Omics Data Including KEGG PLANT Resource

Toshiaki Tokimatsu, Masaaki Kotera, Susumu Goto, and Minoru Kanehisa

Abstract With the rise of experimental technologies for omics research in recent years, considerable quantitative data related to transcription, protein and metabolism are available for predicting protein functions. To predict protein functions from large omics data, reference knowledge databases and bioinformatics tools play considerable roles. KEGG (http://www.genome.jp/kegg/) database we have been establishing is an integrated database of biological systems including genomic, chemical and systemic functional information. Our group has also been developing the tools for genome or chemical analysis as GenomeNet Bioinformatics Tools (http://www.genome.jp/en/gn_tools.html). In this chapter, we introduce the KEGG database resources and the GenomeNet Bioinformatics Tools for predicting protein functions from the viewpoint of omics research, as well as some recent topics (KEGG PLANT Resource and PathPred). KEGG PLANT Resource is one of the contents in the KEGG EDRUG database, and contains links for plant secondary metabolite biosynthesis pathways, plant genomes and EST sequences, chemical information of plant natural products and the prediction tool for plant secondary metabolism pathway. PathPred is a recently developed pathway prediction tool based on the chemical structure transformation patterns of enzyme reactions found in metabolic pathways.

Introduction

In recent years, high-throughput omics data such as transcriptome and metabolome data is continuously increasing. Genomics, transcriptomics and proteomics provide the data of genes and proteins in individual organisms. On the other hand, metabolomics, glycomics, and lipidomics provide information for endogenous molecules, and chemical genomics provides information for exogenous molecules.

T. Tokimatsu (✉)
Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan
e-mail: tokimatu@kuicr.kyoto-u.ac.jp

D. Kihara (ed.), *Protein Function Prediction for Omics Era*,
DOI 10.1007/978-94-007-0881-5_14, © Springer Science+Business Media B.V. 2011

For environmental studies, metagenomics and meta-metabolomics data are becoming available as genomic and chemical information, respectively. One of the main objectives of these high-throughput experiment projects is to uncover molecular building blocks of life. Integration and of high throughput genomics and chemical spaces data and interpretation of high-order function is a powerful technique for understanding molecular building block of life such as protein function. Bioinformatics approaches are required to predict protein function by analyzing exclusively increasing omics data.

KEGG (http://www.kegg.jp/) [1] is a computer representation of biological systems, consisting of a number of sub-databases, such as those including genomic and chemical information. Among these, systems information database is the most unique feature in KEGG. They have been manually collected from review and original articles, other publications, specialists' website, and other resources. In KEGG project, several useful bioinformatics tools have also been developed for genome analysis and chemical analysis. These tools are released as GenomeNet bioinformatics tools at GenomeNet website (http://www.genome.jp/).

Plants are known to produce vast and diverse secondary metabolites, and the total number of plant metabolites are estimated to be over 200,000 [2]. Plant secondary metabolites support our life either directly or indirectly as foods, medicines, and industrial materials. Notably, physiologically active natural products mainly from plants are used as crude drugs and traditional medicine in our lives since ancient times. Physiological active plant natural products have been main resources for drug seed compounds. Thus, elucidating the biosynthetic pathways of plant secondary metabolites is a valuable research area for plant biotechnology, agricultural sciences and pharmaceutical sciences. In just the past decade, transcriptome [3, 4] and metabolome [5, 6] analysis of model plant species such as Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*) has become an active area of research. Accordingly, the necessity of high-quality database resource of plants gains the importance for predicting plant secondary metabolite biosynthesis pathway and protein functions involved in the pathway. Therefore, we are currently accumulating crude drugs and other plant natural product information as KEGG EDRUG database.

In this chapter, we introduce the overview of the KEGG database and the recent topics on the KEGG and GenomeNet resources from the viewpoint of protein function prediction, including KEGG EGENES, KEGG PLANT Resource and PathPred: an enzyme-catalyzed metabolic pathway prediction server.

Outline of KEGG Resource

Overview of KEGG Database

Table 1 shows the list of KEGG main databases [1], and their contents. As of July 2010, KEGG comprises 19 main databases, categorized into systems information, genomic information and chemical information as shown in Table 1. Genomic and

Table 1 KEGG databases

Category	Database	Contents
Systems information	KEGG PATHWAY	Pathway maps
	KEGG BRITE	Functional hierarchies
	KEGG MODULE	Pathway modules
	KEGG DISEASE	Human diseases
	KEGG DRUG	Drugs
	KEGG EDRUG	Crude drugs and other natural products
Genomic information	KEGG ORTHOLOGY	KEGG Orthology (KO) groups
	KEGG GENOME	KEGG organism
	KEGG GENES	Genes in completely sequenced genomes
	KEGG SSDB	Best hit relation within GENES
	KEGG DGENES	Genes in draft genomes
	KEGG EGENES	Genes as EST contigs
	KEGG MGENES	Genes in metagenomes
Chemical information (KEGG LIGAND)	KEGG COMPOUND	Metabolite and other small molecules
	KEGG GLYCAN	Glycans
	KEGG REACTION	Biochemical reactions
	KEGG RPAIR	Reactant pair chemical transformations
	KEGG RCLASS	Reaction classification
	KEGG ENZYME	Enzyme nomenclature

chemical information databases are collection of molecular building blocks of life in the genomic and chemical spaces, respectively, and systems information represent the molecular systems that are built from the molecular building blocks.

KEGG is a computer representation of biological systems. Systems information is the most characteristic feature in KEGG database, and is manually collected from review articles, other publications, specialists' website, and other resources. Six databases in KEGG describe systems information. They are classified into two types. The former three databases (PATHWAY, BRITE, and MODULE) are the databases for pathway and functional classification. The latter three databases (DISEASE, DRUG, and EDRUG) are the databases for analysis of the molecular network-disease association. DISEASE, DRUG, and EDRUG contain data of disease, drug and bioactive natural products, respectively. The following seven databases (ORTHOLOGY, GENOME, GENES, SSDB, DGENES, EGENES, and MGENES) are categorized as genomic information. They are gene catalogs in the completely sequenced genomes, manually defined ortholog groups, computationally calculated sequence similarity information, and supplementary gene catalog data (for draft genomes, EST contigs, and metagenomes). The six databases in chemical information category (COMPOUND, GLYCAN, REACTION, RPAIR, RCLASS, and ENZYME) are collectively called as KEGG LIGAND. They contain the information of small molecules, glycans, biochemical reaction of these molecules, chemical structure transformation patterns derived from reaction data, reaction classification according to the chemical structure transformation pattern, and supplemental information of enzyme nomenclatures. SSDB, DGENES, EGENES, and

MGENES in the genomic information category are computationally generated, but all other 15 databases are manually curated.

KEGG Orthology (KO): Basis of Genome Annotation in KEGG

KEGG Orthology (KO) is the basis for the protein function annotation in KEGG. KEGG ortholog annotation procedure is described as follows. Protein sequences with experimental evidences in specific organisms are used as seeds, and the homologous sequences from other organisms are automatically collected. Consequently, these sequence groups are manually curated and defined as the KO groups in the context of molecular networks; i.e., as the nodes in the KEGG PATHWAY and BRITE. KO groups are given K numbers for identification. Next, cross-species annotation is added as follows. Gene catalogs of all complete genomes are generated from RefSeq database and other public resources. They are computationally processed to generate what we refer to as the GFIT tables, containing the list of genes in a genome with the data of the best-hit genes (i.e., the most homologous genes) against the all other genomes. The automatic cross-species annotation is performed for a set of the "safe" K numbers, representing clearly defined ortholog groups. Manual curation of this automatic annotation is performed using the KOALA and GFIT tools. As of July 2010, genes data taken from 1135 prokaryotes and 131 eukaryotes species are stored in the GENES database. We developed KEGG Automatic Annotation Server (KAAS) as functional annotation tool of genes. This system automatically assigned KO for query genes. Detailed information about KAAS is described in section "KAAS – KEGG Automatic Annotation Server".

PATHWAY and BRITE: Systems Representation in KEGG

KEGG PATHWAY maps describe the dual aspects of metabolic network. The first aspect is genomic information network, i.e., the network of enzyme genes or enzymes. In KEGG PATHWAY, genes and proteins are identified by the K numbers as mentioned in the previous section. EC numbers are shown as the node names in the pathway maps, but they are not used as the identifiers in KEGG. The second aspect is chemical information network, the network of small molecules (chemical compounds) and chemical structure transformations. Chemical compounds are identified by the C numbers and reactions are identified by the R numbers.

The KEGG reference pathway maps and BRITE reference hierarchies are created as to be applicable to all organisms; the exceptions are those describing human diseases. The organism-specific pathways and hierarchies can be generated by using the K numbers as the gene identifiers in particular organisms. Genes in an organism, take *Arabidopsis thaliana* as an example, are annotated with the K numbers, representing manually defined ortholog groups corresponding to the nodes in the KEGG pathway maps.

PATHWAY also provides the global metabolism maps, which are created by manually combining about 120 existing traditional metabolic pathway maps. Circular nodes represent chemical compounds, and the lines connecting two nodes are series of reactions. These global pathway maps allow users to view and compare the entire metabolism, by such means as mapping transcriptome data and/or metabolome data.

Color Objects in KEGG Pathways and BRITE Hierarchies

Integrating large-scale data of genomic (e.g., transcriptome) and/or chemical (e.g., metabolome) spaces onto the systems space (e.g., KEGG PATHWAY, BRITE) helps our understanding for protein function prediction. This section explains the methods for mapping molecular datasets to the KEGG pathway and BRITE hierarchies.

The first method is to use the options "Pathway Mapping" and "Brite Mapping" available on the web pages (http://www.genome.jp/kegg/tool/color_pathway.html and http://www.genome.jp/kegg/tool/color_brite.html), respectively. From the Search Object page, the user can find the objects (genes, metabolites, etc.) of interest in the PATHWAY maps or the BRITE hierarchies by coloring them. Consequently, the user can obtain PATHWAY maps or BRITE hierarchies with these objects favorably colored through the Color Object page. The objects of interest have to be specified by the KEGG identifiers. The user can input the list of objects either directly from the input box or by uploading the file including the list.

Another method for mapping dataset on pathway is accessing KEGG through KEGG API (http://www.genome.jp/kegg/soap/). KEGG API is a web service to use the KEGG system from your program via SOAP/WSDL. The service enables users to develop software that accesses and manipulates vast amount of KEGG data that are constantly updated. KEGG API provides function for coloring pathways. For the general information on KEGG API, please refer to the KEGG API page at GenomeNet (http://www.genome.jp/kegg/soap/).

KEGG REACTION: Chemical Structure Transformation Information in KEGG

KEGG REACTION database contains enzyme reactions taken from KEGG ENZYME database and from the metabolic pathway maps in KEGG PATHWAY database. Each reaction is identified by the R number. KEGG RPAIR database is a collection of reactant pair defined for each reaction in KEGG reaction, together with the chemical structure transformation patterns characterized by the RDM patterns. Each reaction pair is identified by the RP number. In general, a reaction consists of multiple reactant pairs, and the one that appears on the KEGG metabolic pathway maps is called as the main pair. The RDM pattern is defined as KEGG atom type change at the reaction center "R", the difference atom next to reaction center "D", and the matched atom next to reaction center "M", respectively. KEGG RCLASS database represents classification of reaction based on the RDM patterns of main reactant pairs. The transformation pattern may consist of multiple RDM patterns. Each reaction class is identified as RC number. We developed PathPred and E-zyme for predicting pathway and enzymatic functions. The RDM patterns are the basis of these prediction tools. Detailed information about PathPred and E-zyme are described in sections "PathPred: Pathway Prediction Server" and "E-zyme for Prediction of Enzymatic Reactions", respectively.

KEGG Resources and GenomeNet Bioinformatics Tools for Predicting Protein Function

KEGG EDRUG and KEGG PLANT Resource

Overview of KEGG EDRUG and KEGG PLANT Resource

Natural resources including bioactive natural products, such as crude drugs and foods, have been used usefully since ancient times. These natural resources are mostly taken from plants. For this reason, we developed new database, KEGG EDRUG (http://www.genome.jp/kegg/drug/edrug.html), which is a database of crude drugs, essential oils and other useful natural product resources including plant information resources.

Plants are known to produce diverse chemical compounds including those with medicinal, nutritional and industrial values. These plant secondary metabolites can be divided into groups that share the same core substructure, originated from the same biosynthetic pathways and chemical building blocks. In this context, KEGG EDRUG is also considered as a part of KEGG PLANT Resource, which is an interface to the KEGG resource for plant research, especially for understanding relationships between genomic and chemical information of plant natural products. KEGG PLANT Resource links to the biosynthetic pathway of plant secondary metabolites, sequences of plant genomes and ESTs, structural classification of plant secondary metabolites, and pathway prediction tools for plant secondary metabolites.

Pathway Maps of Plant Secondary Metabolite Biosynthesis

Plants produce vast and diverse secondary metabolites, but core structures of these metabolites are synthesized from several important precursors. Biosynthesis pathways of plant secondary metabolites are classifiable by precursors and their biosynthesis pathways. Grouping secondary metabolite biosynthesis pathways by their biosynthetic origins and mapping the core structures on overview pathway maps will help our understanding about plant secondary metabolisms.

Therefore, the KEGG PLANT Resource provides the links to the pathway maps for secondary metabolism. There are three types of pathway maps in KEGG PLANT Resource, e.g., KEGG traditional pathway maps, a global map, and overview maps.

In recent years, the repertoire of the KEGG pathway maps for plant secondary metabolism is expanded, and some of the maps are renewed incorporating recent information. As a result, the secondary metabolite biosynthesis subclass in KEGG PATHWAY is divided to “Metabolism of Terpenoids and Polyketides” and “Biosynthesis of Other Secondary Metabolites”.

Recently, we developed new global map of secondary metabolism pathways. Figure 1a is a screenshot of the reference global maps of secondary metabolite

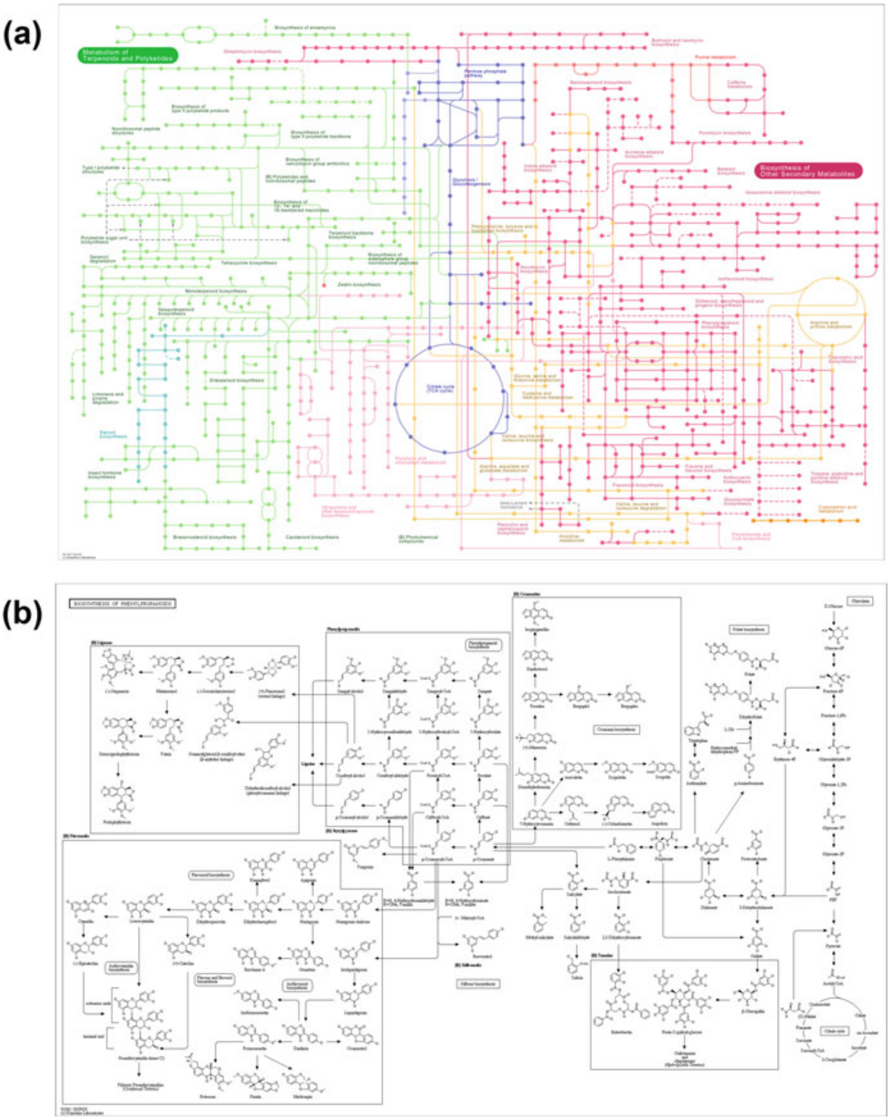

(a)
Metabolism of Terpenoids and Polyketides
Biosynthesis of Other Secondary Metabolites
Citrate cycle (TCA cycle)
(b)
BIOSYNTHESIS OF PHENYLPROPANOIDS

Fig. 1 Example PATHWAY maps of global pathway and overview pathway. (**a**) Reference global pathway map of secondary metabolites. (**b**) Overview pathway map of Biosynthesis of phenylpropanoids

biosynthesis, which include secondary metabolite biosynthesis pathways and related pathway maps. This map allows users to view and compare the omics data or mapping species-specific pathways on the secondary metabolite specific pathway. The major part of this pathway is related to plant metabolism, intended for the usefulness for plant scientists. This new global map of secondary metabolism also allows users to map the species-specific pathways by using K numbers, e.g. *Arabidopsis thaliana*. Users can easily figure out the secondary metabolism of specific species at a glance.

KEGG PLANT Resource also provides two levels of overview pathway maps for plant secondary metabolite biosynthesis. First level map is general overview pathway map for biosynthesis of secondary metabolites. This pathway map is based on overview of biosynthetic pathways map in KEGG PATHWAY, and modification includes plant-specific pathways such as biosynthesis of secondary metabolite core structures. Core structures of major plant secondary metabolites are mapped on the overview map. The second are the category maps of plant secondary metabolite pathways starting from specific precursor biosynthesis pathway. The category maps include the detailed information of the core structure biosynthesis pathway. The category maps reflect the classification of plant secondary metabolites. Figure 1b is the example screenshot of overview pathway map of secondary metabolite biosynthesis. Different to standard KEGG pathway maps, these overview maps contain the graphics of chemical structures and not for mapping species-specific information or experimental data.

KEGG GENES and EGENES of Plants: Sequence Information of Plant Species

Although some research groups have launched genome projects of model plants for specific plant families and industrial important plants, the available complete genomes for plants are still very limited in comparison to other organism groups such as animals and bacteria. At the end of June 2010, only 13 plant species of complete genomes have been published and stored in KEGG, including two draft genomes (Table 2). Thus, massive EST dataset have been processed for a number of plant species to generate the EGENES database where the EST contigs are treated as genes and automatically annotated with the KO (K number) identifiers [7] by KAAS automatic annotation (see section "KAAS – KEGG Automatic Annotation Server"). Currently, 77 plant species of the EST datasets are stored in the EGENES database. The EST dataset covers wider variety of plant families and species than complete genomes, especially in asterids as shown in Table 2.

We also provide the BRITE hierarchical lists of plant phylogenetic classification. The phylogenetic classification of angiosperm is based on the second Angiosperm Phylogeny Group classification for the orders and families of flowering plants (APG II 2003) [8]. APG II classification is based on molecular systematic of flowering plants.

Table 2 Number of plant families and species in complete genomes and EST datasets in KEGG

Classification	Complete genomes (GENES, DGENES)	EST datasets (EGENES)
Eudicots: asterids	0(0)	7(18)
Eudicots: resides	4(4)	8(31)
Eudicots: others	1(1)	4(4)
Monocots	1(3)	3(11)
Basal angiosperms	0(0)	1(1)
Gymnosperms	0(0)	2(5)
Ferns	0(0)	1(2)
Mosses	1(1)	2(2)
Green algae	2(3)	2(2)
Red algae	1(1)	0(0)
Glaucophytes	0(0)	1(1)
Total	10(13)	31(77)

Figures before parentheses are the number of families and figures in parenthesis are number of species

Classification of Plant Secondary Metabolites

Based on the biosynthetic origin, major plant secondary metabolites are classified to polyketides (from acetate-malonate pathway), phenylprpoanoids and related compounds (from shikimate pathway), terpenoids and steroids (from mevalonate pathway or deoxyxylulose-phosphate pathway), and nitorogen-containing alkaloids and sulfur-containing compounds (from amino-acids and related compounds). This classification also reflects the core chemical structures of plant secondary metabolites. We classified the plant secondary metabolites in the KEGG COMPOUND database by their biosynthetic origins and the core chemical structures. Currently, about 2600 plant secondary metabolites are collected in the BRITE classification of phytochemical compounds (Table 3). The top seven metabolite classes contain over 150 metabolites. All these metabolite classes are well studied, and contain many bioactive metabolites such as components of crude drugs and essential oils. This classification system is also used for categorizing drugs and other bioactive compounds derived from plant metabolites. In a future, phytochemical compound classification will be more refined and categorized into detail core structures, which will help to link compound classification to biosynthetic pathways.

KEGG EDRUG Database

KEGG EDRUG database is a collection of bioactive natural products, which are ingested in our bodies. These natural products, such as crude drugs, essential oils, etc., are mostly supplied from plant. The E number is used for identifier of each KEGG EDRUG entry, and is associated with the chemical components, efficacy information, and source species information. At present, crude drugs and essential oils are the two main components of the EDRUG entries.

Table 3 Phytochemical compounds classification in KEGG BRITE as of July 2010

Classification	# of compounds
Total phytochemical compounds	2617
Phenylpropanoids and related compounds	744
Monolignols	37
Lignans	71
Coumarins	61
Flavonoids	507
Stilbenoids	39
Hydrolysable tannins	24
Misc. Phenylpropanoids	5
Polyketides	97
Quinones	41
gamma-Pyrones	56
Terpenoids	965
Hemiterpenoids (C5)	3
Monoterpenoids (C10)	164
Sesquiterpenoids (C15)	298
Diterpenoids (C20)	170
Triterpenoids (C30), sterols and steroids	286
Tetraterpenoids (C40) (Carotenods)	36
Polyterpenoids	8
Alkaloids	720
Alkaloids derived from ornithine	103
Alkaloids derived from lysine	76
Alkaloids derived from nicotinic acid	18
Alkaloids derived from tyrosine	201
Alkaloids derived from tryptophan and anthranillic acid	199
Alkaloids derived from histidine	3
Alkaloids derived by amination reactions	103
Misc. alkaloids	17
Amino acid derivatives other than alkaloids	91
Betalains	30
Cyanogenic glucosides	25
Glucosinolates	36

This table shows the first classes and second classes of the Phytochemical compounds classification in KEGG BRITE

Pathway Prediction for Plant Secondary Metabolism

Predicting biosynthetic pathways of plant secondary metabolites and linking them to the plant genomes are challenging problems. We recently developed a web-based server named PathPred, which is designed to predict secondary metabolite biosynthesis pathways for a given compound using the information of known enzyme reactions (i.e., RDM patterns and chemical structure alignments of substrate-product pairs). Detailed information about PathPred is described in the next section.

PathPred: Pathway Prediction Server

PathPred (http://www.genome.jp/tools/pathpred/) [9] is a recently developed web-based server for predicting metabolic pathway of a given compound. Current version of PathPred provides a multi-step reaction prediction of xenobiotics biodegradation pathways and secondary metabolite biosynthesis pathways, and we aim to improve this toward more sophisticated prediction for metabolic pathway reconstruction.

Prediction procedure consists of the following three steps. The first step is a global similarity search of a query compound against the KEGG COMPOUND database by the SIMCOMP program [10, 11]. The second step is a local pattern match against the RDM pattern library to select the matched patterns that are applicable to the query compound. Specific category of the KEGG pathways, such as xenobiotics biodegradation pathways or secondary metabolite biosynthesis pathways, have their specific subsets of the RDM patterns [9, 12, 13]. Thus, we extracted and use the specific RDM patterns library for xenobiotics biodegradation and secondary metabolite biosynthesis, respectively. The third step is to apply the structure transformation to the query compound based on the selected matched patterns. PathPred has a function to assign plausible EC numbers to the suggested reaction steps. This function is based on the E-zyme program. Further information about E-zyme is described in section "E-zyme for Prediction of Enzymatic Reactions ".

We have to mention that the meaning of the query compound is different depending on whether users would like to predict biodegradation pathways or biosynthesis pathways. In the case of biodegradation, the query compound is the molecule that will break down. In other words, it is located at the beginning of the pathway. On the contrary, in the case of biosynthesis, the query compound is the molecule that was synthesized. In other words, it is located at the end of the pathway. This makes sense when we consider what we would like to do, but this is sometimes confusing when we actually use this application. Users may optionally input the end product in biodegradation or the start compound in the biosynthesis. If the users have an idea what the origin of the query secondary metabolite might be, specifying them might help better prediction. Users can use input query compound in the MDL mol format, the SMILES representation, or KEGG COMPOUND/DRUG identifier (C/D numbers). We provide KegDraw for drawing chemical compound structures and glycan structures. Compound structures drawn by KegDraw are also used as queries for PathPred. KegDraw is java application and software for MacOSX, Microsoft Windows, and Linux provided from KegTools download page (http://www.genome.jp/download/).

Figure 2 shows an example of the prediction result for plant secondary metabolite pathway (shown as a tree-like structure) by PathPred. This example includes the biosynthetic pathway from umbelliferon (7-Hydroxycoumarin) to fraxidin (8-Hydroxy-6,7-dimethoxycoumarin). Fraxidin is a major component of a crude drug Saposhnikovia root [14] and chemically classified to coumarins. Figure 2a shows a PathPred prediction pathway tree from umbelliferon to fraxidin. Red, blue, and gray compound numbers indicate the query or final compounds, compounds in the KEGG database, and hypothetical molecules that are also generated elsewhere in

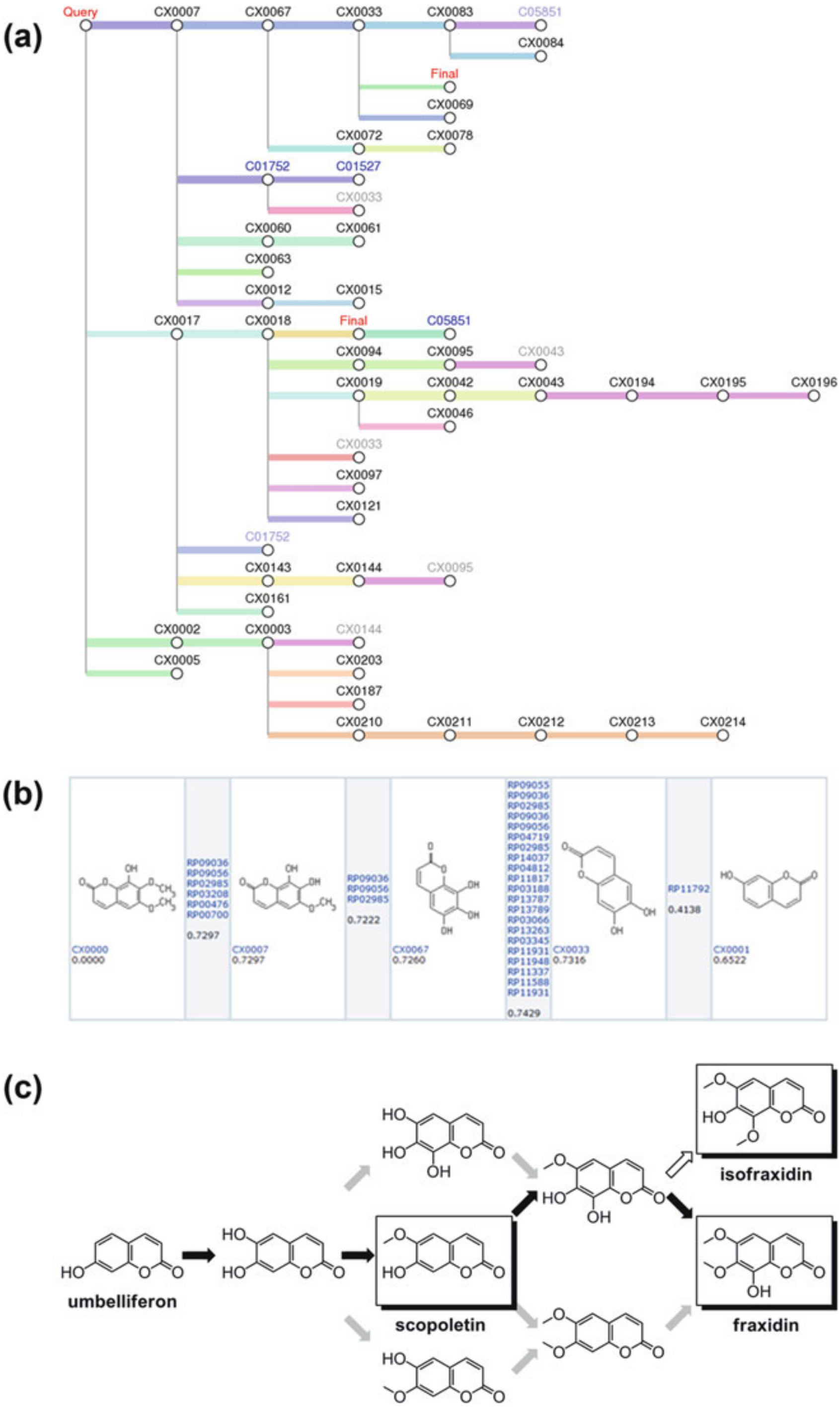


Fig. 2 (continued)

the tree, respectively. In the case of secondary metabolite biosynthesis pathway prediction, “query” and “final” mean “end product” and “starting substrate” of the synthetic pathway, respectively. Figure 2b shows one of the predicted paths from query (fraxidin) to final (umbelliferon) compounds in the prediction pathway tree. The identification numbers (which we refer to as the RP numbers) between two chemical structures indicate the links to the template reaction pairs for predicting the transformation. Figure 2c shows the total predicted pathway network from umbelliferon to fraxidin and 7-hydroxycoumarin related components of crude drug Saposhinicovia root. Black and gray arrows are the paths predicted by PathPred. Three Saposhinicovia components, fraxidin, isofraxidin, and scopoletin [14], are located on or linked to the predicted pathway. According to the components of Saposhinicovia root, the predicted path indicated by the black arrows pathway is highly likely to exist (black arrows pathway is same as Fig. 2b).

As PathPred is knowledge-based prediction system, the quality of the knowledgebase is crucial for the prediction accuracy. We are continuously updating the KEGG RPAIR, REACTION and PATHWAY databases. We also categorized the plant secondary metabolite biosynthesis pathway into subclasses, such as phenylpropanoids, polyketides, terpenoids and alkaloids, to use only the frequent RDM patterns depending on the compound subclasses and to improve the efficiency in terms of specificity and computational time.

E-zyme for Prediction of Enzymatic Reactions

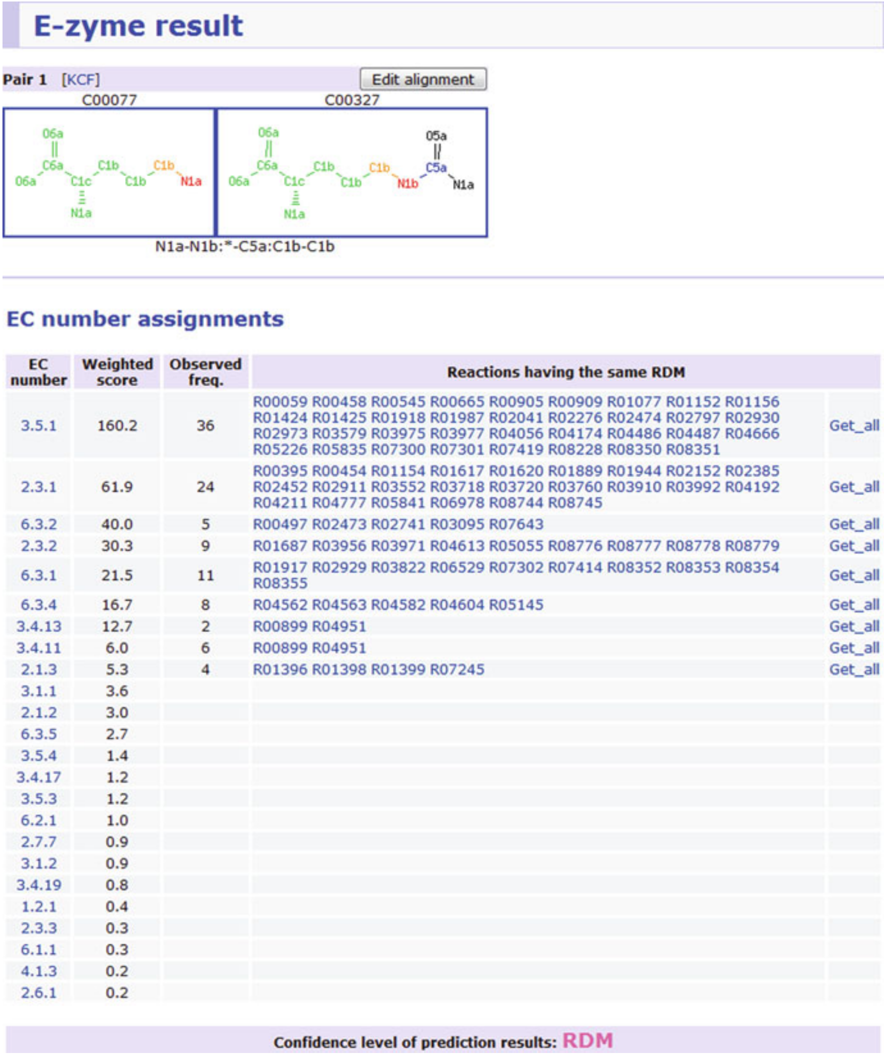
Enzyme Commission (EC) number [15, 16] is a hierarchical classification system for enzyme reactions established by International Union of Biochemistry and Molecular Biology (IUBMB). This EC number system is widely accepted as the standard classification system in the field of biochemical and enzymatic studies. The EC numbers also play key roles in linking the enzyme genes or proteins to reactions and in the computational representations of enzymatic reactions in metabolic pathways. E-zyme (http://www.genome.jp/tools/e-zyme/) [12, 17] is the GenomeNet bioinformatics tool for prediction of enzyme reactions, i.e., to automatically assign the EC numbers up to the sub-sub classes for a given enzyme reaction. The prediction process is based on the relationships between the EC numbers and the corresponding RDM patterns (See section “Color Objects in KEGG Pathways and BRITE Hierarchies”).

◀

Fig. 2 Prediction result of biosynthetic pathway from umbelliferon (7-Hydroxycoumarin) to fraxidin (8-Hydroxy-6,7-dimethoxycoumarin) by PathPred. (**a**) Predicted pathway tree consist compounds (*node*) and reactions (*edge*). (**b**) One of successfully predicted pathway from umbelliferon to fraxidin. (**c**) Possible predicted pathway from umbelliferon to fraxidin by PathPred (*black* and *gray arrows*) and Saposinicovia coumarin components (in *boxes*). *Brack arrowed* pathway is same pathway as shown in figure (**b**)

Prediction of EC classes by E-zyme consists of the following steps. First, the chemical structures of a substrate and a product are compared by the SIMCOMP chemical structure alignment program [10, 11], and outputs the changes occurred during the reaction in a form of the RDM patterns. Consequently, the possible EC numbers are suggested based on the pre-computed correlations between the RDM patterns and the EC numbers. Users can input query compounds in the MDL mol format, or KEGG COMPOUND identifier (C number).

Figure 3 shows an example screenshot of the output by E-zyme. It includes the chemical alignment of the two compound structures (i.e., a substrate and a product



E-zyme result

Pair 1 [KCF] Edit alignment

C00077 C00327

N1a-N1b:*-C5a:C1b-C1b

EC number assignments

EC number	Weighted score	Observed freq.	Reactions having the same RDM	
3.5.1	160.2	36	R00059 R00458 R00545 R00665 R00905 R00909 R01077 R01152 R01156 R01424 R01425 R01918 R01987 R02041 R02276 R02474 R02797 R02930 R02973 R03579 R03975 R03977 R04056 R04174 R04486 R04487 R04666 R05226 R05835 R07300 R07301 R07419 R08228 R08350 R08351	Get_all
2.3.1	61.9	24	R00395 R00454 R01154 R01617 R01620 R01889 R01944 R02152 R02385 R02452 R02911 R03552 R03718 R03720 R03760 R03910 R03992 R04192 R04211 R04777 R05841 R06978 R08744 R08745	Get_all
6.3.2	40.0	5	R00497 R02473 R02741 R03095 R07643	Get_all
2.3.2	30.3	9	R01687 R03956 R03971 R04613 R05055 R08776 R08777 R08778 R08779	Get_all
6.3.1	21.5	11	R01917 R02929 R03822 R06529 R07302 R07414 R08352 R08353 R08354 R08355	Get_all
6.3.4	16.7	8	R04562 R04563 R04582 R04604 R05145	Get_all
3.4.13	12.7	2	R00899 R04951	Get_all
3.4.11	6.0	6	R00899 R04951	Get_all
2.1.3	5.3	4	R01396 R01398 R01399 R07245	Get_all
3.1.1	3.6			
2.1.2	3.0			
6.3.5	2.7			
3.5.4	1.4			
3.4.17	1.2			
3.5.3	1.2			
6.2.1	1.0			
2.7.7	0.9			
3.1.2	0.9			
3.4.19	0.8			
1.2.1	0.4			
2.3.3	0.3			
6.1.1	0.3			
4.1.3	0.2			
2.6.1	0.2			

Confidence level of prediction results: RDM

Fig. 3 Example screenshot of E-zyme out put page of the prediction result

of a possible reaction) and the list of the predicted EC numbers. The alignment of the compounds and assigned RDM patterns are shown in the upper section of the result page. In the lower section of the result page, EC number prediction results are displayed.

For further information, we recommend to refer Yamanishi et al. [17] and documents in E-zyme page (http://www.genome.jp/tools/e-zyme/).

KAAS – KEGG Automatic Annotation Server

In recent years, the number of complete and draft genomes, EST and metagenome sequences are rapidly increasing. This makes it increasingly important to automatically annotate functional properties and biological roles to genes. We provide KAAS (http://www.genome.jp/tools/kaas/) [18] for this purpose as a GenomeNet bioinformatics tool. KAAS serves functional annotation of genes in genomes (or large number of genes) by BLAST comparisons against the manually curated KEGG GENES database. Genes in KEGG DGENES (draft genomes), KEGG EGENES (EST contigs) and KEGG MGENES (metagenomes) are automatically annotated by KAAS.

We also provide a web service for the general public users. Overall procedure of the KAAS annotation is as follows. KAAS accepts three types of query sets, i.e., complete or draft genome, partial genome, or EST sequences. Query sequence data should be in multi-FASTA format of amino-acid or nucleotide sequences with unique ids. The user can choose one or more species from the latest KEGG GENES entries as the reference data set. We recommend to choose more closely related species to the species of interest as possible, in order to obtain better result.

KAAS provides three types of outputs as the results. “KO list” is the flat list of the correspondence table with query genes and K numbers assigned by the KAAS program. “BRITE hierarchies” is the hierarchical list of the annotated genes, which is incorporated into the classification of the BRITE database. “Pathway map” is the list of pathways that include the annotated query genes. The list is linked to the graphical pathway maps, and the annotated query genes are highlighted.

For further information, refer to Moriya et al. [18] and the documents in the KAAS webpage (http://www.genome.jp/tools/kaas/).

KegArray

To predict protein function from omics data, one of the effective ways is the integrated analysis of omics data by using systems information such as pathway network diagram. KegArray is a standalone desktop application for analyzing both transcriptome data (gene expression profiles) and metabolome data (compound profiles) in conjunction with the KEGG databases (http://www.genome.jp/download/kegtools.html). KegArray software is a Java application, and users can download the software

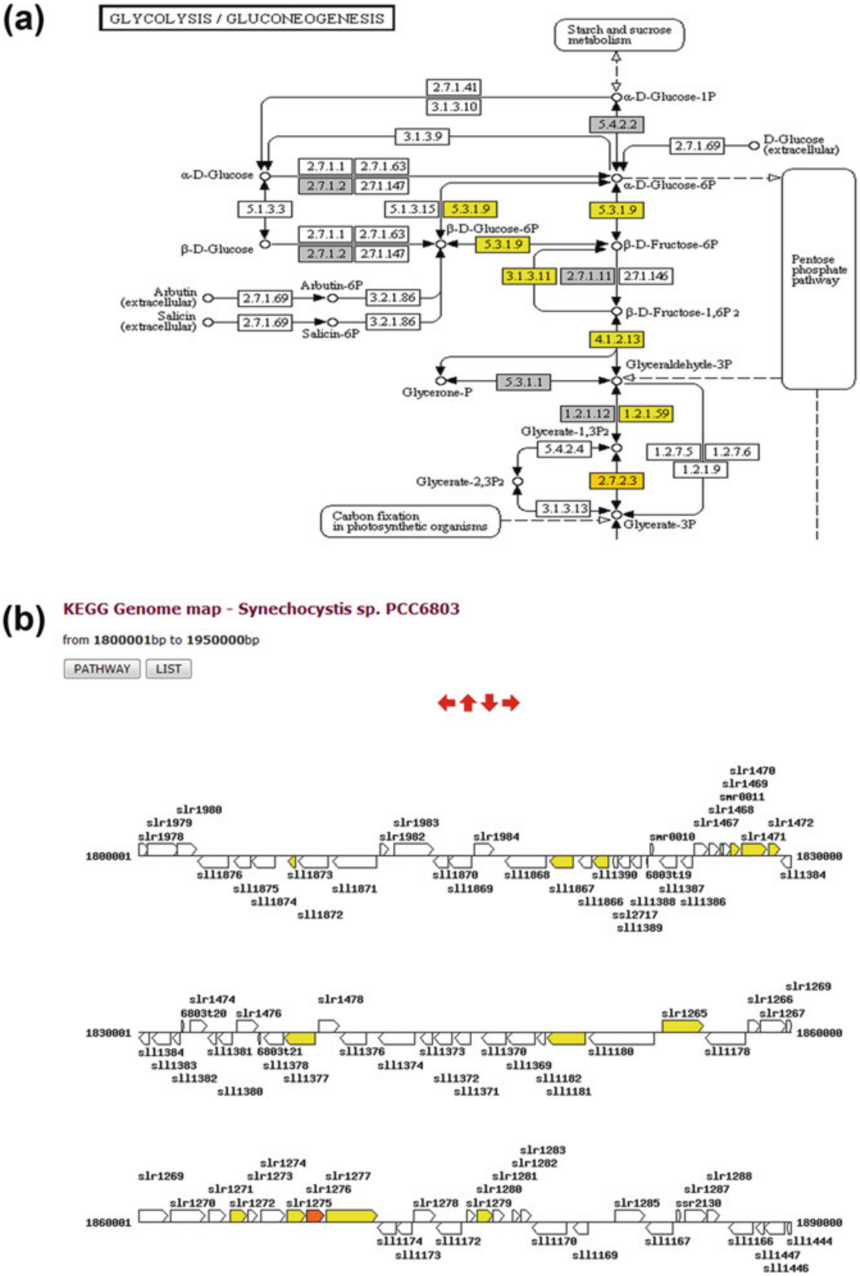

(a)
GLYCOLYSIS / GLUCONEOGENESIS
Starch and sucrose metabolism
2.7.1.41
3.1.3.10
α-D-Glucose-1P
5.4.2.2
3.1.3.9
2.7.1.69
D-Glucose (extracellular)
α-D-Glucose
2.7.1.1
2.7.1.63
2.7.1.2
2.7.1.147
α-D-Glucose-6P
5.1.3.3
5.1.3.15
5.3.1.9
β-D-Glucose-6P
β-D-Glucose
β-D-Fructose-6P
Pentose phosphate pathway
Arbutin (extracellular)
Arbutin-6P
3.2.1.86
Salicin (extracellular)
Salicin-6P
3.1.3.11
2.7.1.11
2.7.1.146
β-D-Fructose-1,6P2
4.1.2.13
Glyceraldehyde-3P
Glycerone-P
5.3.1.1
1.2.1.12
1.2.1.59
Glycerate-1,3P2
5.4.2.4
1.2.7.5
1.2.7.6
1.2.1.9
Glycerate-2,3P2
2.7.2.3
3.1.3.13
Carbon fixation in photosynthetic organisms
Glycerate-3P
(b)
KEGG Genome map - Synechocystis sp. PCC6803
from 1800001bp to 1950000bp
PATHWAY
LIST
1800001
1830000
slr1978 slr1979 slr1980 slr1982 slr1983 slr1984 smr0010 slr1467 slr1468 smr0011 slr1469 slr1470 slr1471 slr1472
sll1876 sll1875 sll1874 sll1873 sll1872 sll1871 sll1870 sll1869 sll1868 sll1867 sll1866 sll1390 sll1388 ssl2717 sll1389 6803t19 sll1387 sll1386 sll1384
1830001
1860000
slr1474 6803t20 slr1476 slr1478 slr1265 slr1266 slr1267 slr1269
sll1384 sll1383 sll1382 sll1381 sll1380 6803t21 sll1378 sll1377 sll1376 sll1374 sll1373 sll1372 sll1371 sll1370 sll1369 sll1182 sll1181 sll1180 sll1178
1860001
1890000
slr1269 slr1270 slr1271 slr1272 slr1273 slr1274 slr1275 slr1276 slr1277 slr1278 slr1279 slr1280 slr1281 slr1282 slr1283 slr1285 ssr2130 slr1287 slr1288
sll1174 sll1173 sll1172 sll1170 sll1169 sll1167 sll1166 sll1447 sll1446 sll1444

Fig. 4 Example screenshot of (**a**) pathway mapping and (**b**) genome mapping by KegArray

(for Mac OS X, Microsoft Windows, and Linux) from the download page (http://www.genome.jp/download/).

Transcriptome data format for KegArray is KEGG EXPRESSION format or tab-deliminated text similar to the KEGG EXPRESSION format. KEGG EXPRESSION format is original data format for KEGG EXPRESSION database (http://www.genome.jp/kegg/expression/). KEGG EXPRESSION database is a repository of microarray gene expression profile data for *Synechosystis*, *Bacillus subtilis* and other species. KegArray can convert external database IDs (e.g. NCBI GI) to the KEGG GENES IDs. Only ratio values can be used for metabolome data. Main function of KegArray is to map the transcriptome and metabolome data to the KEGG resources including PATHWAY, BRITE and genome maps. Figure 4 shows example screen shots of pathway mapping and genome mapping by KegArray tools. As shown in the figure, users can visualize the up- or down-regulated genes on various KEGG systems information. Users can also visualize increasing or decreasing metabolites on various KEGG objects.

Detailed usage information for KegArray, refer to the ReadMe file provided in the KegTools download page (http://www.genome.jp/download/).

Summary

In the post-genomic era, bioinformatics approach is necessary to analyze increasing omics data. Also, high quality database and bioinformatics approach will play important roles to predict protein function. KEGG is a manually curated integrated database for computer representation of biological systems. KEGG and GenomeNet also provide several useful tools to support protein function prediction from omics data.

In this chapter, we briefly outlined a perspective of the KEGG database and several tools for predicting protein functions from omics data, with introducing some recent topics. KEGG EDRUG is the database for crude drugs, essential oils, other natural products and related plant resources. KEGG EDRUG provides useful information for protein function related to plant secondary metabolism pathway. PathPred and E-zyme are tools for predicting enzyme reaction pathway from metabolites. These tools help users to predict unknown pathway. KAAS is automatic annotation and pathway prediction server for large set of sequences. KegArray is desktop application for transcriptome and metabolome analysis. KegArray will help mapping those data to the KEGG systems information such as KEGG PATHWAY, KEGG BRITE etc.

Acknowledgements The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University. The KEGG project is supported by the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency, and a grant-in-aid for scientific research on the priority area “Comprehensive Genomics” from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Kanehisa, M., Goto, S., Furumichi, M., et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. **38**: D355–D360 (2010).
2. Dixon, R.A., Strack, D. Phytochemistry meets genome analysis, and beyond. Phytochemistry **62**: 815–816 (2003).
3. Donson, J., Fang, Y., Espiritu-Santo, G., et al. Comprehensive gene expression analysis by transcript profiling. Plant Mol. Biol. **48**: 75–97 (2002).
4. Aharoni, A., Vorst, O. DNA microarrays for functional plant genomics. Plant Mol. Biol. **48**: 99–118 (2002).
5. Sumner, L.W., Mendes, P., Dixon, R.A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry **62**: 817–836 (2003).
6. Sato, S., Soga, T., Nishioka, T., et al. Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. Plant J **40**: 151–163 (2004).
7. Masoudi-Nejad, A., Goto, S., Jauregui, R., et al. EGENES: transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG. Plant Physiol. **144**: 857–866 (2007).
8. The Angiosperm Phylogeny Group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot. J. Linnean Soc. **141**: 399–436 (2003).
9. Moriya, Y., Shigemizu, D., Hattori, M., et al. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acid Res. **38**: W138–W143 (2010).
10. Hattori, M., Okuno, Y., Goto, S., et al. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J. Am. Chem. Soc. **125**: 11853–11865 (2003).
11. Hattori, M., Tanaka, N., Kanehisa, M., et al. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. Nucleic Acid Res. **38**: W652–W656 (2010).
12. Kotera, M., Okuno, Y., Hattori, M., et al. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. J. Am. Chem. Soc. **126**: 16487–16498 (2004).
13. Oh, M., Yamada, T., Hattori, M., et al. Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. J. Chem. Inf. Model. **47**: 1702–1712 (2007).
14. Okuyama, E., Hasegawa, T., Matsushita, T., et al. Analgestic Components of Saposhinikovia Root (*Saposhinikovia divaricata*). Chem. Pharm. Bull. **49**: 154–160 (2001).
15. Barrett, A.J., Cantor, C.R., Liebecq, C., et al. *Enzyme nomenclature*. San Diego, CA: Academic (1992).
16. Tipton, K.F., Boyce, S. History of the enzyme nomenclature system. Bioinformatics **16**: 34–40 (2000).
17. Yamanishi, Y., Hattori, M., Kotera, M., et al. E-zyme: prediction potential EC numbers from the chemical transformation pattern of substrate-product pairs. Bioinformatics **25**: i79–i86 (2009).
18. Moriya, Y., Itoh, M., Okuda, S., et al. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acid Res. **35**: W182–W185 (2007).