Systematic Analysis of Enzyme-Catalyzed Reaction Patterns and Prediction of Microbial Biodegradation Pathways

Mina Oh, Takuji Yamada, Masahiro Hattori, Susumu Goto, and Minoru Kanehisa*

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Received January 9, 2007

The roles of chemical compounds in biological systems are now systematically analyzed by high-throughput experimental technologies. To automate the processing and interpretation of large-scale data it is necessary to develop bioinformatics methods to extract information from the chemical structures of these small molecules by considering the interactions and reactions involving proteins and other biological macromolecules. Here we focus on metabolic compounds and present a knowledge-based approach for understanding reactivity and metabolic fate in enzyme-catalyzed reactions in a given organism or group. We first constructed the KEGG RPAIR database containing chemical structure alignments and structure transformation patterns, called RDM patterns, for 7091 reactant pairs (substrate-product pairs) in 5734 known enzyme-catalyzed reactions. A total of 2205 RDM patterns were then categorized based on the KEGG PATHWAY database. The majority of RDM patterns were uniquely or preferentially found in specific classes of pathways, although some RDM patterns, such as those involving phosphorylation, were ubiquitous. The xenobiotics biodegradation pathways contained the most distinct RDM patterns, and we developed a scheme for predicting bacterial biodegradation pathways given chemical structures of, for example, environmental compounds.

INTRODUCTION

In comparison with the biological macromolecules of nucleic acids and proteins, chemical compounds have long been considered a minor class of molecules in the biological sciences. This situation is rapidly changing due to the increasing activity of metabolomics, chemical genomics, and chemical biology in which the biological functions of a large number of small molecules are uncovered at the molecular, cellular, and organism levels.[1−3] In order to best utilize the large-scale data sets being generated, bioinformatics methods have to be developed to extract the information encoded in the small molecular structures and to understand the information in the context of molecular interactions and reactions involving proteins and other biomolecules.

We have been developing both the database resource[3] and computational methods[4,5] for integrated analysis of genomic (protein and nucleic acid) and chemical (small molecule) information. Here small molecules are divided into two categories. One is metabolic compounds that are subject to enzyme-catalyzed reactions that maintain the biological system. The other category is regulatory compounds that interact with proteins, DNA, RNA, and other endogenous molecules to regulate or perturb the biological system. Thus far, we have focused on metabolic compounds. Our aim has been to understand the universe of enzyme-catalyzed reactions, which would hopefully be represented by higher-level knowledge for use in further analyses and predictions. In our previous work we extracted chemical transformation patterns in enzymatic reactions for automating the EC number assignment[5] based on the atom typing and the chemical structure comparison method.[4] In this paper we report the construction of the KEGG RPAIR database, a compilation of manually curated chemical structure transformation patterns in enzymatic reactions, and we discuss an application using this database.

We apply the database to predict the metabolic fate of small molecules in the living cell, which is an important problem for the interpretation of metabolomics and other experimental data, drug discovery, and evaluation of an organism's response to environmental challenges.[6] In particular, we focus on the biodegradation pathways of xenobiotics in bacteria. Since these compounds or their metabolic byproducts can be potentially toxic to mammals, the investigation of xenobiotics biodegradation pathways will have practical importance in human health and environmental remediation.[7]

First, we perform a systematic survey of the KEGG RPAIR database (http://www.genome.jp/kegg/reaction/) containing chemical structure alignments of substrate-product pairs (reactant pairs) and chemical structure transformation patterns in all known enzyme-catalyzed reactions, both of which are computationally generated and manually curated. Because multiple substrates and products are usually involved in an enzymatic reaction, only biochemically meaningful pairs are defined according to the six EC classes of enzymes.[5] Biochemical structure transformations are described by what we call RDM patterns, which represent KEGG atom type changes at the reaction center atom (R atom) and its neighboring atoms on the different (mismatched) region (D atom) and the matched region (M atom). Thus, the RPAIR database is a library of RDM patterns, representing our current knowledge on the universe of enzyme-catalyzed reactions. The library also provides a generalization of complex reactions in cellular metabolism, enabling prediction of potential reaction centers and structure transformations

* Corresponding author e-mail: kanehisa@kuicr.kyoto-u.ac.jp.

10.1021/ci700006f CCC: $37.00
Published on Web 05/22/2007

that could happen, given new chemical compound structures. The prediction power will be improved by examining relationships between specific RDM patterns and specific metabolic pathway categories in the KEGG PATHWAY database. It would further be improved by combining additional knowledge, such as the tendency to form consecutive reaction steps and corresponding gene sets (e.g., operons) in the genome.

Second, we present a method to predict potential biodegradation pathways of xenobiotics. There have been previous prediction methods such as METEOR, META, MetabolExpert, and PPS in UM-BBD. METEOR is a stand-alone Windows program for prediction of mammalian xenobiotics metabolism;[8] META covers aerobic, anaerobic biodegradation, and photodegradation;[9,10] MetabolExpert includes animal metabolism, plant metabolism, photodegradation chemistry, and soil degradation chemistry;[11] and the PPS in UM-BBD predicts microbial biodegradation pathways.[12,13] These methods are similar in that they use a rule-based approach, recognizing functional groups and their transformation patterns in organic compounds. A limitation is uncertainty on whether the rules are sufficient for diverse compounds and diverse pathways.

In contrast, our approach is based on data, rather than rules. Current knowledge on enzyme-catalyzed reactions is stored in the library of RDM patterns, which is continuously updated as new reactions are added to the KEGG system.[3] The RDM patterns presumably represent the reaction specificity of enzymes but not the substrate specificity. Therefore, in our prediction system a new compound is first compared against all known substrate and product structures, and then possible RDM patterns are selected by considering the similarity scores of matched compounds. By limiting the data set to bacterial reactions appearing in the “Xenobiotics Biodegradation and Metabolism” category of the KEGG PATHWAY database, this prediction system can be adjusted to microbial biodegradations.

MATERIALS AND METHODS

Databases of Reactions and Reactant Pairs. The KEGG REACTION database (http://www.genome.jp/kegg/reaction/) contains all known enzyme-catalyzed reactions taken from the IUBMB Enzyme Nomenclature[14] (http://www.chem-.qmul.ac.uk/iubmb/enzyme/) and also from the metabolic pathway section of the KEGG PATHWAY database. In KEGG release 40.0 (October 2006), there were 6807 reactions, including 3222 IUBMB reactions. An enzyme-catalyzed reaction usually involves multiple substrates and products. A reaction is decomposed into a set of substrate-product pairs, and each of the binary pairs with common structural moieties is defined as a reactant pair. A more specific definition of reactant pairs in each of the six enzyme classes is given in our previous work.[5] In the present analysis, we used 5734 reactions, including 3009 IUBMB reactions, because the other reactions lacked well-defined chemical structures for substrates and/or products, such as for glycans and other biopolymers. Thus, 7091 reactant pairs were extracted and stored in the KEGG RPAIR database.

Chemical Structure Comparison. Two chemical structures of each reactant pair are aligned to obtain the common substructure by applying the chemical structure comparison method named SIMCOMP.[4] This method is based on a graph matching algorithm for finding the maximum common subgraph in two graphs where the node of the graph (chemical compound) is labeled by KEGG atom types. This atom typing distinguishes functional groups and atomic environments, resulting in 68 atom types: 23 for carbon, 16 for nitrogen, 18 for oxygen, 7 for sulfur, 2 for phosphorus, and 1 each for halogens and other undefined atoms.[4] Generally KEGG atom types are represented with three characters/numbers such as C1a, N2b, and O7x. The first character indicates the atomic species, the second number represents the information on the kind of atomic bonding, and the third character represents the information on the kind and the number of substituted groups.

The optimal alignment generated by SIMCOMP is subjected to manual curation, because the program may terminate the search for large structures and repetitive structures and report a suboptimal solution due to CPU time constraints.

Definition of RDM Pattern. The optimal alignment generated by SIMCOMP and verified by human inspection is then used to define a chemical structure transformation pattern of a reactant pair by examining the matched and nonmatched regions. The boundary atom between the matched and nonmatched regions of a reactant pair is defined as a reaction center or the R (Reaction center) atom. The atom(s) adjacent to the R atom in the nonmatched region is(are) defined as the D (Difference region) atom(s), and the atom(s) adjacent to the R atom in the matched region is(are) defined as the M (Matched region) atom(s).[5] The KEGG atom type changes that occur at the R, D, and M atoms during enzyme-catalyzed reaction constitute an RDM pattern, which characterizes the chemical structure transformation of a reactant pair. An example of an RDM pattern is shown in Figure 1.

In most cases an RDM pattern consists of a change in a single atomic region (subgraph of R, D, and M atoms), represented as

$$\text{R1-R2}:\ \text{D1}-\text{D2}:\ \text{M1}-\text{M2}$$

However, it may consist of multiple regions, in which case an RDM pattern is represented as a collection of these atomic region changes.

Pathway-Based Categorization of RDM Patterns. The RDM patterns are classified according to the categorization of metabolic pathways in the KEGG PATHWAY database (http://www.genome.jp/kegg/pathway.html). The Metabolism section of this database consists of 11 categories: Carbohydrate Metabolism (abbreviated in this paper as: Carbohydrate), Energy Metabolism (Energy), Lipid Metabolism (Lipid), Nucleotide Metabolism (Nucleotide), Amino Acid Metabolism (Amino acid), Metabolism of Other Amino Acids (Other amino acids), Glycan Biosynthesis and Metabolism (Glycan), Biosynthesis of Polyketides and Nonribosomal Peptides (PK/NRP), Metabolism of Cofactors and Vitamins (Cofactor/vitamin), Biosynthesis of Secondary Metabolites (Other secondary metabolites), and Xenobiotics Biodegradation and Metabolism (Xenobiotics).

In order to examine any characteristic distribution of RDM patterns in these categories, PermutMatrix software[15,16] was used to perform a cluster analysis and to display the results graphically. The similarity of RDM patterns is based on the

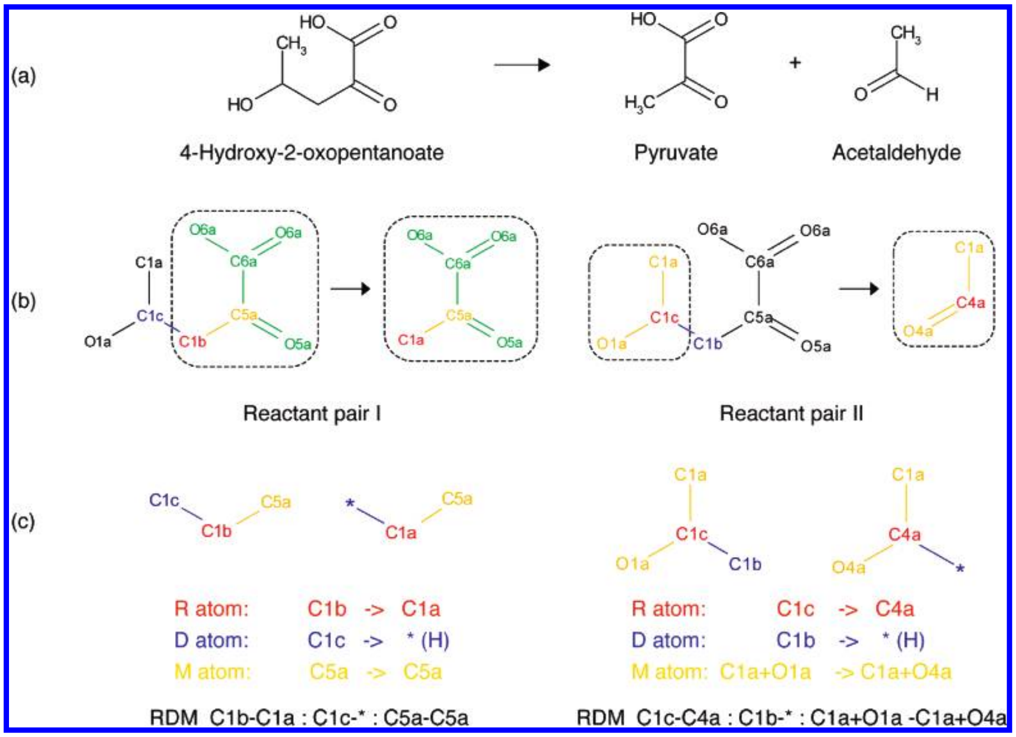


Figure 1. The alignment of the reactant pair and the definition of RDM in an enzyme-catalyzed reaction (R00750 in KEGG). (a) The overall reaction where 4-hydroxy-2-oxopentanoate (C03589) is catalyzed to pyruvate (C00022) and acetaldehyde (C00084) by lyases (aldehyde-lyases or oxo-acid-lyases: EC 4.1.2.- or 4.1.3.-). (b) The reaction is decomposed into a couple of reactant pairs: reactant pair I (A01083) containing pyruvate and reactant pair II (A01084) containing acetaldehyde. The matched substructures obtained by SIMCOMP alignments are shown by dotted boxes for both pairs. Each structure is labeled with the KEGG atom types, and the RDM atoms are colored with red, blue, and yellow, respectively. The matched structure except the M atoms is colored green. (c) The RDM patterns are extracted from the two reactant pairs, where asterisks indicate hydrogen atoms. The RDM pattern is a set of KEGG atom type changes, such as C1b-C1a in the R atom, C1c-* in the D atoms, and C5a-C5a in the M atoms for reactant pair I.

frequency in each category and measured by the Euclidean distance. The Euclidean distance between two points $P = (p_x)$ and $Q = (q_x)$ is defined as

$$\sqrt{\sum_x (p_x - q_x)^2} = |P - Q|$$

where p_x and q_x are the frequencies of RDM patterns in the metabolic pathway category x.

RESULTS

Statistics of RDM Patterns. We constructed the KEGG RPAIR database, which contains 7091 reactant pairs (substrate-product pairs) involving 4302 compounds derived from 5734 enzyme-catalyzed reactions. Each reactant pair is associated with an RDM pattern for the chemical transformation pattern, and there are 2205 different RDM patterns in the RPAIR database (Table 1). The majority of RDM patterns, 1406 out of 2205 (64%), appeared only once in the RPAIR database, while one pattern appeared 175 times, which was the largest. When the number of RDM patterns was plotted against the number of occurrences in the log−log scale, the relationship was linear (data not shown). Note that an RDM pattern defined here may contain multiple atomic region changes, which means that a reactant pair may have multiple reaction centers (see Materials and Methods). The majority of RDM patterns, 1499 out of 2205 (68%), consisted of single atomic regions (a single reaction center), and 706 patterns contained up to nine regions (multiple reaction centers).

Table 1. Statistics of the KEGG RPAIR Database[a]

number of reactant pairs	7091
number of reactions	5734
number of compounds	4302
number of RDM patterns	2205
number of RD types	1407
number of RM types	1406
number of R atom types	610
number of D atom types	854
number of M atom types	1093

[a] As of December 11, 2006.

A total of 7091 reactant pairs were assigned from 5734 reactions involving 4302 compounds. Out of 2205 RDM patterns in the 7091 reactant pairs, the numbers are shown here for the different combination types of RD and RM and also for the different atom types of R, D, and M.

Unique and Frequently Observed RDM Patterns in Specific Pathway Categories. The compiled RDM patterns relate to specific biological processes represented by the KEGG pathway database. We first examined if any RDM patterns were unique to specific categories of KEGG metabolic pathways. We used a subset of 2205 patterns corresponding to the reactions appearing in the KEGG metabolic pathways. At the same time, we distinguished the direction of the reaction, thereby double-counting the same pattern. The KEGG pathway map provides information on whether a reaction is reversible or an irreversible, and this can be parsed from the KGML (KEGG markup language) file. In the case of a reversible reaction, the RDM patterns of the forward and reverse reactions were considered

Table 2. Number and Uniqueness of RDM Patterns in Each Category of KEGG Metabolic Pathways[a]

metabolic pathway category	reaction	compd	RDM	unique RDM	uniqueness (%)
1. carbohydrate	727	452	444	303	68.2
2. energy	171	114	128	62	48.4
3. lipid	560	450	261	169	64.8
4. nucleotide	256	143	102	66	64.7
5. amino acid	723	534	430	254	59.7
6. other amino acids	170	147	119	55	46.2
7. glycan	243	96	42	14	33.3
8. PK/NRP	210	226	97	66	68.0
9. cofactor/vitamin	336	279	224	151	67.4
10. other secondary metabolites	533	514	245	161	65.7
11. xenobiotics	580	522	347	275	79.3
total	4256	3057	1901	1576	

[a] The numbers of reactions, compounds, and RDM patterns were categorized according to the 11 categories of the KEGG metabolic pathways. The number of unique RDM patterns in each category was then obtained and the uniqueness (%) was calculated.

separately. Thus, we extracted 1901 RDM patterns from 4256 reactions in the KEGG metabolic pathways. Table 2 shows the classification and uniqueness of these RDM patterns in the 11 KEGG pathway categories. The number of unique patterns was highest in the pathways for xenobiotics biodegradation in which roughly 80% of the RDM patterns were exclusive to this category. This suggests the possibility of using these RDM patterns for biodegradation pathway prediction.

Table 2 indicates that there are 325 nonunique RDM patterns that appear in multiple metabolic pathway categories. It also indicates that 72 nonunique patterns are found in the xenobiotics category. We then asked if there is a tendency of these patterns to appear more often in specific categories. Figure 2 shows the matrix of 97 frequently occurring, nonunique RDM patterns versus 11 categories, which was obtained by the PermutMatrix clustering program. It graphically displays clusters of highly observed patterns by coloring cells on the basis of measured frequencies. Here we used the threshold value of 5 for the minimum frequency of occurrence in at least one category. The red color intensity of each matrix cell represents the frequency, with black representing values below the threshold and the brightest red for the maximum of 57. The clusters identified are labeled (a)−(e) on the left, containing RDM patterns that appear in (a) multiple categories, which we call general metabolic pathways, (b) amino acid metabolism, (c) xenobiotics biodegradation, (d) carbohydrate metabolism, and (e) lipid metabolism. The numbers of RDM patterns constituting clusters (a)−(e) were 5, 33, 10, 17, and 9, respectively. Thus, the (c) cluster for xenobiotics biodegradation consisted of 10 RDM patterns, and this cluster always remained significant even when the threshold value was lowered to zero. Because the clusters (b)−(e) cover most of the RDM patterns, it can be concluded that the majority of frequently observed RDM patterns are almost exclusively found in specific metabolic pathway categories.

Examples of General and Characteristic RDM Patterns. An interesting observation in the pathway-based categorization of the RDM patterns is that the majority of RDM patterns are characteristic of specific metabolic pathway categories. However, there are some RDM patterns that are distributed throughout all the metabolic pathways. Two outstanding RDM patterns in cluster (a) of Figure 2 are as follows: O1a - O2b : * - P1b : C1b - C1b and Olc - O2c : * - P1b : P1b - P1b. They are related to phosphotransferases (kinases) catalyzing phosphorylations and phosphohydrolases (phosphatases) catalyzing dephosphorylations (Figure 3(a)). Here, ATP or orthophosphate is used as a phosphate donor group, and the reaction involves the transfer of a phosphate group to a substrate (for example, a ribonucleotide phosphate), yielding a phosphorylated product and ADP. Another interesting point is that these general RDM patterns are frequently observed in most metabolic pathways except for xenobiotics biodegradation. We therefore assume that the reactions related to xenobiotics biodegradation have more specific reaction patterns than other metabolic pathway reactions.

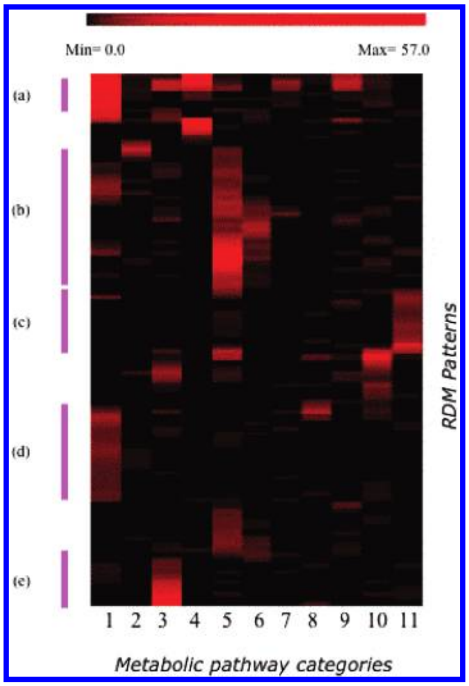
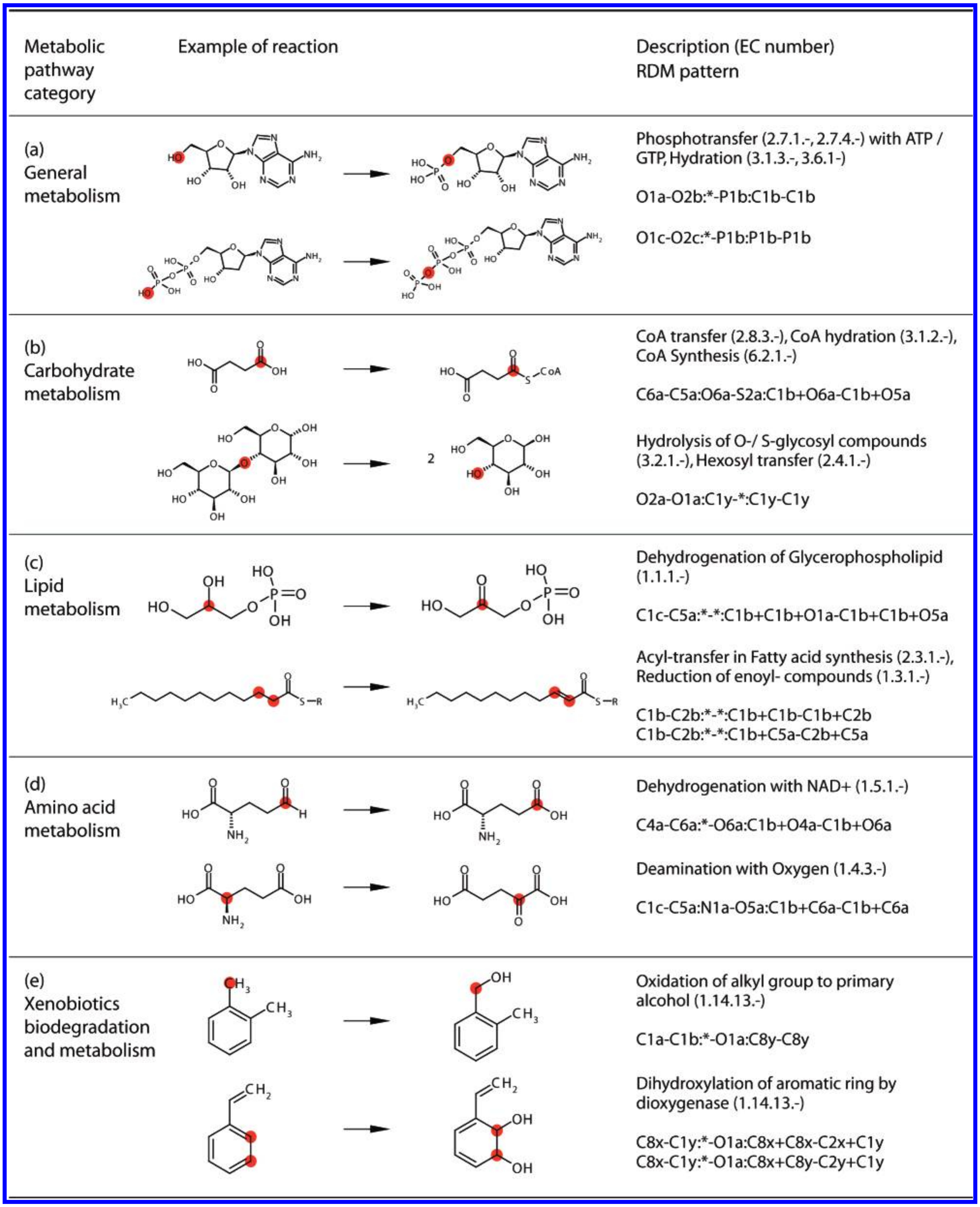


Figure 2. Clustering of 97 frequently occurring, nonunique RDM patterns according to the 11 metabolic pathway categories. Each RDM pattern is represented by a single row; each metabolic pathway category is represented by a single column, where the rightmost column (labeled 11) corresponds to xenobiotics degradation (labels 1−11 follow the metabolic pathway category of Table 2). The red color intensity of each matrix cell represents the frequency, black meaning below the threshold and the brightest red for 57. Five clusters are indicated by the bars labeled with (a), (b), (c), (d), and (e). The clusters (b), (c), (d), and (e) appear mostly in the fifth, 11th, first, and third columns for amino acid metabolism, xenobiotics biodegradation, carbohydrate metabolism, and lipid metabolism, respectively. Cluster (a) appears in most categories.

In Figure 2, characteristic RDM patterns are clustered in carbohydrate metabolism, lipid metabolism, amino acid metabolism, and xenobiotics biodegradation. We identified those frequently observed nonunique patterns, called preferential patterns, in each pathway category with the frequency threshold of 5. We also identified frequently observed unique patterns in Table 2 (see the Supporting Information) with the frequency threshold of 5. Some examples of these RDM patterns and representative reactions are shown in Figure 3(b)−(e).

For carbohydrate metabolism 26 unique patterns (out of 303 in Table 2) and 17 preferential patterns (Figure 2) were



Metabolic pathway category	Example of reaction	Description (EC number) RDM pattern
(a) General metabolism		Phosphotransfer (2.7.1.-, 2.7.4.-) with ATP / GTP, Hydration (3.1.3.-, 3.6.1-) O1a-O2b:*-P1b:C1b-C1b O1c-O2c:*-P1b:P1b-P1b
(b) Carbohydrate metabolism		CoA transfer (2.8.3.-), CoA hydration (3.1.2.-), CoA Synthesis (6.2.1.-) C6a-C5a:O6a-S2a:C1b+O6a-C1b+O5a Hydrolysis of O-/ S-glycosyl compounds (3.2.1.-), Hexosyl transfer (2.4.1.-) O2a-O1a:C1y-*:C1y-C1y
(c) Lipid metabolism		Dehydrogenation of Glycerophospholipid (1.1.1.-) C1c-C5a:*-*:C1b+C1b+O1a-C1b+C1b+O5a Acyl-transfer in Fatty acid synthesis (2.3.1.-), Reduction of enoyl- compounds (1.3.1.-) C1b-C2b:*-*:C1b+C1b-C1b+C2b C1b-C2b:*-*:C1b+C5a-C2b+C5a
(d) Amino acid metabolism		Dehydrogenation with NAD+ (1.5.1.-) C4a-C6a:*-O6a:C1b+O4a-C1b+O6a Deamination with Oxygen (1.4.3.-) C1c-C5a:N1a-O5a:C1b+C6a-C1b+C6a
(e) Xenobiotics biodegradation and metabolism		Oxidation of alkyl group to primary alcohol (1.14.13.-) C1a-C1b:*-O1a:C8y-C8y Dihydroxylation of aromatic ring by dioxygenase (1.14.13.-) C8x-C1y:*-O1a:C8x+C8x-C2x+C1y C8x-C1y:*-O1a:C8x+C8y-C2y+C1y

Figure 3. Examples of general (a) and characteristic (b)−(e) RDM patterns and representative reactions. The reaction center (R) atoms are colored with red circles in the reactions.

identified with the frequency threshold of 5. The vast majority of these patterns comprise reactions involving CoA ligases (ADP forming) like succinyl-CoA synthetase. It is obvious that these reactions related to ATP or coenzyme A are central to carbohydrate metabolism. There are also characteristic RDM patterns for reversible reactions converting between the cyclic form of sugars and the open chain form. Another characteristic pattern is the hydrolyase reaction for pentose or hexose to form a disaccharide from two monosaccharides by a glycosidic bond. Still others include many reactions for oxidation of a hydroxyl group of pentose or hexose to a ketone group or carboxylic acid by oxidoreductase with EC 1.1.1.-, 1.1.2.-, 1.1.3.-, 1.1.1.5.-, 1.1.99.-, 1.2.1.-, and 1.2.99.-.

For lipid metabolism 18 (9 unique and 9 preferential) characteristic RDM patterns were found, and examples of

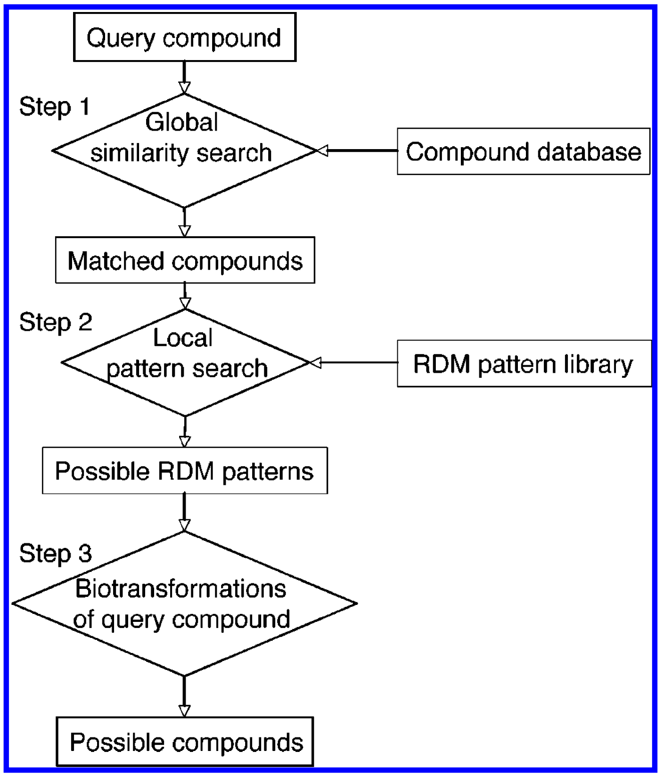


Figure 4. A flow chart of the prediction scheme for enzyme-catalyzed reaction pathways. The first step is global similarity searching of a query compound against the compound database by the SIMCOMP program. The input format of a query compound is KEGG chemical function (KCF) format for representation of KEGG atom types. The compound database is grouped according to the 11 KEGG metabolic pathway categories. Thus, the degradation pathway prediction is performed by using the xenobiotics compound library. The second step is local pattern searching of matched compounds against the RDM pattern library. Finally, the query compound is transformed to possible compounds by following the transformation patterns of matched compounds. The generated compounds are then used as next query and the prediction cycle is repeated.

reactions are as follows. The saturated aliphatic chains next to a ketone group, not only in CoA compounds linked with saturated fatty acids but also in the steroid rings, often eliminate water to form a double bond with EC 2.3.1.85 or 2.3.1.86. Furthermore, the oxidation of the hydroxyl group to the ketone group by a cofactor like NAD+, NADP+, or FAD in phospholipids or the steroid rings is specific to lipid metabolism.

For amino acid metabolism 43 (10 unique and 33 preferential) characteristic RDM patterns were found. The most characteristic reaction is an instantaneous reaction of both deamination and aminotransfer with oxygen. Furthermore, oxidoreduction between aldehyde and carboxylate by N-acetyl transferase with NADH or N-succinyl transferase with acetyl-CoA or succinyl-CoA appear characteristically in amino acid metabolism.

For xenobiotics degradation 16 (6 unique and 10 preferential) characteristic RDM patterns were found. One of the two most frequently occurring RDM patterns is the reaction in which an aromatic aldehyde is transformed to an aromatic acid with a cofactor such as NAD+ by dehydrogenase (EC 1.2.1.-, 1.2.3.-). In the other reaction an alkyl group attached to aromatic ring is easily oxidized to yield a primary alcohol by oxidoreductase with EC 1.14.13.-. In addition, an outstanding reaction is dioxygenation of aromatic rings (or aliphatic groups) using oxygen and NAD(P)H as a catalyst. The dioxygenation (EC 1.14.13.-) removes aromaticity of aromatic rings, and aromaticity is successively recovered by dehydrogenation with EC 1.3.1.19/56/66.

Table 3. Prediction of UM-BBD Reactions[a]

threshold	correct (a)	incorrect (b)	no hits (c)	accuracy (a)/(a+b)	coverage (a+b)/(a+b+c)
0.9	14 (14)	4 (3)	102	78% (82%)	15%
0.8	23 (27)	19 (17)	78	55% (61%)	35%
0.7	31(47)	51 (35)	38	38% (57%)	68%

[a] The number in parentheses indicates that the correct one was present among multiple choices.

Prediction Scheme for Xenobiotics Degradation. The analysis of RDM patterns presented above clearly indicates that the reactions in xenobiotics biodegradation pathways are distinct from the other reactions. We attempt here to predict biodegradation pathways of a query compound (such as new xenobiotics). In order to predict bacterial biodegradation pathways, we have excluded reactions in the KEGG pathway map for metabolism of xenobiotics by cytochrome P450 (map00980) from the data set, because this map represents mostly mammalian P450 metabolism.

We introduced two types of structure searches. One is a local substructure matching against the known RDM pattern library, which can be used for inferring a converted structure based on the matched RDM transformation pattern. The other is a global structure matching of the query compound and the database compound, which is to prioritize more similar compounds for selection of RDM patterns from many possibilities. These two types of matching presumably represent the reaction specificity and the substrate specificity of an enzyme-catalyzed reaction. There is also an implicit assumption that the variations of substrate recognition would be generated by paralogous genes in bacterial genomes without much affecting the conserved catalytic activity.

Figure 4 illustrates our overall prediction scheme. The first step is to compare the structure of a query compound with those of all known substrates and products in the KEGG xenobiotics category (excluding map00980). We have used the SIMCOMP program and the library of 326 compounds with RDM atom assignments out of 441 compounds currently found in this category. Because a compound can have multiple RDM subgraphs (atom assignments), and because an RDM subgraph can be converted to other subgraphs in multiple ways, this compound library corresponds to 461 RDM transformation patterns. The second step is to select from this RDM pattern library possible RDM patterns based on the matched compounds and the structural similarity scores obtained from SIMCOMP. In the third step, the selected RDM patterns are used to infer how the query compound would be transformed to plausible compounds. This of course is based on the assumption that the query compound will mimic the biotransformation patterns of the best matched compounds. Finally the generated compound is used as the next query compound, and the prediction may be repeated.

Prediction Accuracy. Obviously, the success of our prediction scheme depends on whether similar compounds and similar RDM patterns can be found in the database. To

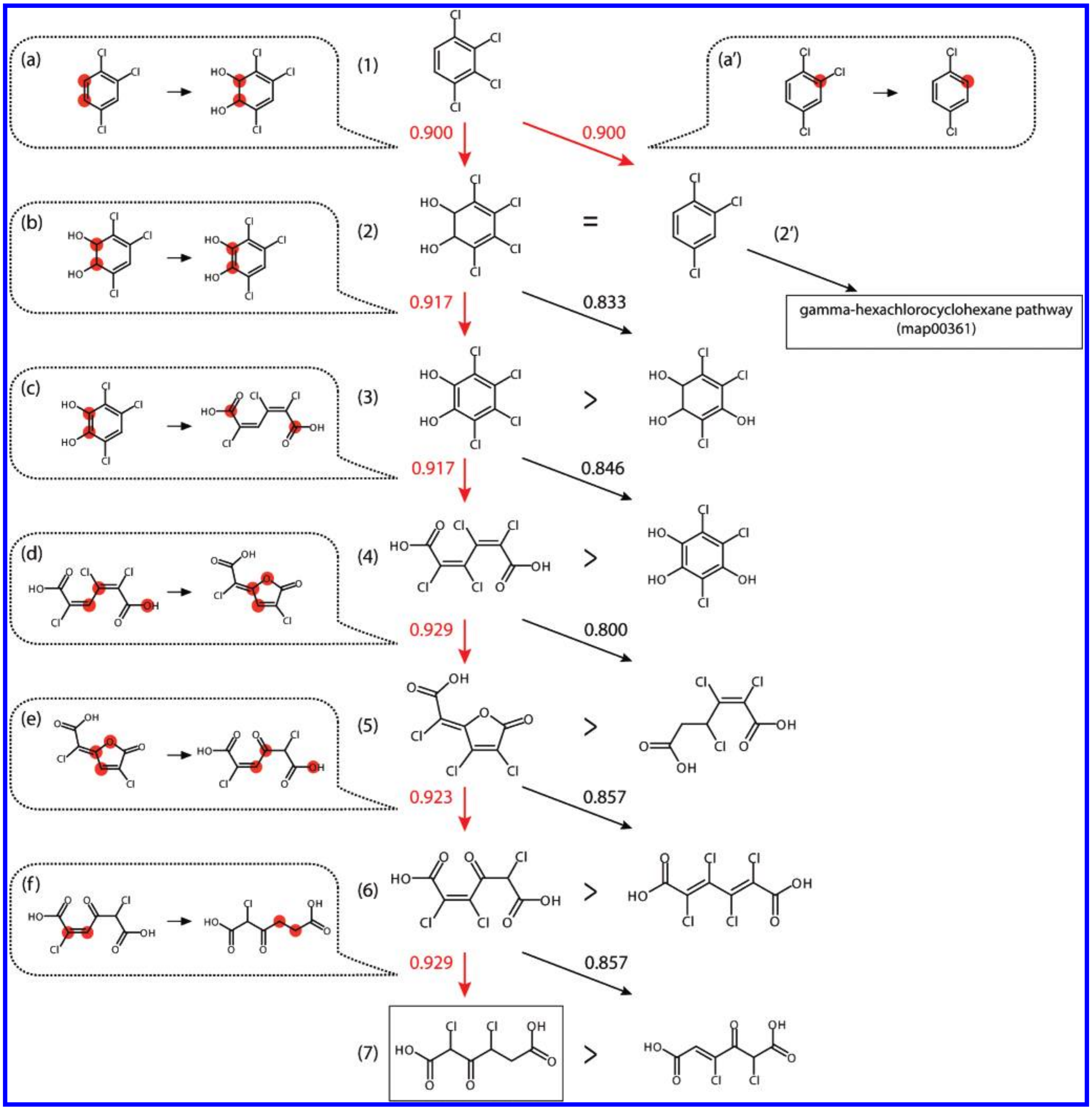


Figure 5. The prediction of 1,2,3,4-tetrachlorobenzene biodegradation pathway. The query compound (1) is transformed to compounds (2)−(7) with the transformation patterns of matching compounds (a)−(f). The compounds are as follows: (1) 1,2,3,4-tetrachlorobenzene, (2) 3,4,5,6-tetrachloro-1,2-dihydroxycyclohexa-3,5-diene, (2′) 1,2,4-trichlorobenzene, (3) 3,4,5,6-tetrachlorocatechol, (4) 2,3,4,5-tetrachloromuconate, (5) 2,4,5-trichlorocarboxymethylenebut-2-en-4-olide, (6) 2,3,5-trichloro-4-oxohex-2-enedioate, (7) 2,4-dichloro-3-oxoadipate, (a) 1,2,4-trichlorobenzene, (a’) 1,2,4-trichlorobenzene, (b) 3,4,6-trichloro-cis-1,2-dihydroxycyclohexa-3,5-diene, (c) 3,4,6-trichlorocatechol, (d) 2,3,5-trichloromuconate, (e) 2,5-dichlorocarboxymethylenebut-2-en-4-olide, and (f) 2,5-dichloro-4-oxohex-2-enedioate. The reactions in the dotted boxes show the transformation patterns of the best matching compounds (a)−(f), where the reaction center atoms are marked with red circles. Just for a comparison purpose, alternative transformation patterns are also shown using the second best matching compounds. Arrows are associated with the similarity scores, and those colored red represent the predicted pathways with the best matching compounds.

estimate prediction accuracy, we used 120 chemical compounds in the UM-BBD database,[12] which are not yet incorporated in KEGG, and compared our prediction with the reaction reported in UM-BBD. The result is shown in Table 3, where the threshold of compound similarity score in SIMCOMP is varied from 0.9 to 0.7. Coverage was not high for higher threshold values, meaning that no similar compounds were found in our data set of 326 compounds. The accuracy was measured in two ways; either (a) the best matched compound with the highest similarity score or (b) any of the matched compounds above threshold was used to obtain the RDM pattern to reproduce the correct reaction in UM-BBD. Although there is still room for improvement in the prediction scheme itself, Table 3 seems to suggest that as we gather more experimentally verified data and expand the library of RDM patterns, the prediction accuracy will increase for the majority of new chemical compounds.

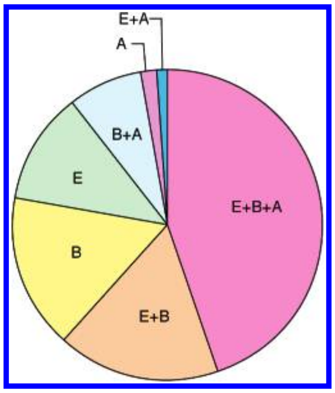


Figure 6. Statistics of the organism group specific RDM patterns. Out of 2205 RDM patterns in our data set, 990 patterns had correspondence with the KEGG Orthology (KO) system. They were classified according to the three organism groups: eukaryotes (E), bacteria (B), and archaea (A). The numbers of RDM patterns in the pie chart are as follows: E: 116 (11.7%), B: 159 (16.1%), A: 17 (1.7%), E+B: 167 (16.9%), B+A: 76 (7.7%), E+A: 10 (1.0%), and E+B+A: 445 (44.9%).

Prediction of Biodegradation Pathway of 1,2,3,4-Tetrachlorobenzene. When a matching compound is found, our prediction worked relatively well. One successful example was the biodegradation pathway for a halogenated xenobiotic, 1,2,3,4-tetrachlorobenzene. This compound is used as a component of dielectric fluids and in organic synthesis and is known to be a highly recalcitrant pollutant. *Pseudomonas chlororaphis* RW71 mineralizes 1,2,3,4-tetrachlorobenzene as a sole source of carbon and energy, thereby releasing stoichiometric amounts of chloride.[17] No data were available for this compound or organism in KEGG. In general, many halogenated compounds are among the most common environmental pollutants,[18] and thus the study on degradation pathways of these compounds is becoming increasingly important.[19]

The result of our prediction is shown in Figure 5. In the first cycle, the query compound 1,2,3,4-tetrachlorobenzene (1) (alphanumerical labels in parentheses correspond to those shown in the figure) was best matched to 1,2,4-trichlorobenzene (a) with the similarity score of 0.9. However, 1,2,3,4-tetrachlorobenzene (1) has two alternatives, because 1,2,4-trichlorobenzene (C06594) has two RDM patterns (A09390 and A09389) in the KEGG RPAIR database. One is oxidation of an aromatic ring by means of dioxygenase (EC 1.14.12.).[20,21] The other is a dehalogenase reaction.[22] By dehalogenation, 1,2,3,4-tetrachlorobenzene (1) would be transformed to 1,2,4-trichlorobenzene (2′). There is a biodegradation pathway of 1,2,4-trichlorobenzene (2′) in the gamma-hexachlorocyclohexane degradation map (map00361) of the KEGG PATHWAY database. Thus, the biodegradation of 1,2,3,4-tetrachlorobenzene (1) could be linked to this pathway map although it is not certain whether an enzyme catalyzing 1,2,3,4-tetrachlorobenzene (1) to 1,2,4-trichlorobenzene (2′) exists or not.

On the other hand, the first possibility of dioxygenase (adding oxygen) will produce 3,4,5,6-tetrachloro-1,2-dihydroxycyclohexa-3,5-diene (2). We then carried out the second cycle of prediction using this compound as a query. We obtained the best matching compound, 3,4,6-trichloro-cis-1,2-dihydroxycyclohexa-3,5-diene (b) (similarity score 0.917). Using the RDM pattern on the matched compound as a template, 3,4,5,6-tetrachloro-1,2-dihydroxycyclohexa-3,5-diene (2) was transformed to 3,4,5,6-tetrachlorocatechol (3) by means of a dehydrogenase system (NAD+ oxidoreductase: EC 1.3.1.-).[23] In the third cycle, 3,4,5,6-tetrachlorocatechol (3) was matched to 3,4,6-trichlorocatechol (c) (similarity score 0.917). With the matched RDM pattern, the 1,2-aromatic bond of 3,4,5,6-tetrachlorocatechol (3) would be broken by dioxygenase (EC 1.13.11.1) and converted to 2,3,4,5-tetrachloromuconate (4).[24] In the fourth cycle, the matching compound, 2,3,5-trichloromuconate (d) (similarity score 0.929), was used to predict that chloromuconate cycloisomerase (EC 5.5.1.7) would eliminate a chloride in 2,3,4,5-tetrachloromuconate (4), yielding 2,4,5-trichlorocarboxymethylenebut-2-en-4-olide (5).[25] In the fifth cycle, this was transformed by a hydrolase (EC 3.1.1.45) to 2,3,5-trichloro-4-oxohex-2-enedioate (6), according to the matching compound 2,5-dichlorocarboxymethylenebut-2-en-4-olide (e) (similarity score 0.923).[26] In the sixth cycle, 2,3,5-trichloro-4-oxohex-2-enedioate (6) was transformed to 2,4-dichloro-3-oxoadipate (7), in accordance that 2,5-dichloro-4-oxohex-2-enedioate (f) (similarity score 0.929) is transformed to 2-chloro-3-oxoadipate by maleylacetate reductase (EC 1.3.1.32).[27] Our predicted pathway from 1,2,3,4-tetrachlorobenzene (1) to 2,4-dichloro-3-oxoadipate (7) was identical to the known 1,2,3,4-tetrachlorobenzene degradation pathway in UM-BBD.

DISCUSSION

Genetic and Biosynthetic Codes. The macromolecular structures of DNA, RNA, and proteins are determined by template-based syntheses of replication, transcription, and translation with the genetic code. In contrast, the structures of glycans, lipids, polyketides, nonribosomal peptides, and various secondary metabolites are determined by biosynthetic pathways. Such biosynthetic codes are far more complex than the genetic code, and our knowledge of them is still quite limited. Bacteria and plants are known to produce diverse substances, many of which have medical and pharmaceutical relevance, including antibiotics and crude drugs. Bacteria are also known to degrade exotic compounds, including environmental compounds that they have never encountered before. With the complete genome sequences available for an increasing number of organisms, it should in principle be possible to infer a complete set of biochemical substances produced by each organism and also to infer a set of xenobiotic substances that can be degraded by an organism.

Toward this end, we have been developing bioinformatics approaches for linking the genomic space of enzyme genes and the chemical space of endogenous substances by organizing our knowledge on biosynthetic and biodegradation pathways. The analysis of carbohydrate sugar chains was the most straightforward because only a limited class of enzymes is involved. The genomic or transcriptomic repertoire of glycosyltransferase genes was linked to possible glycan structures using the KEGG GLYCAN database.[28,29] Similarly, genomic information was used to predict the chemical structures of another class of polymeric chains, polyketides, and nonribosomal peptides.[30] For various other small molecule metabolites, because many different types of enzymes are involved, it is essential to organize and digitize higher-level knowledge about the universe of enzyme-catalyzed reactions. The RDM pattern library is an

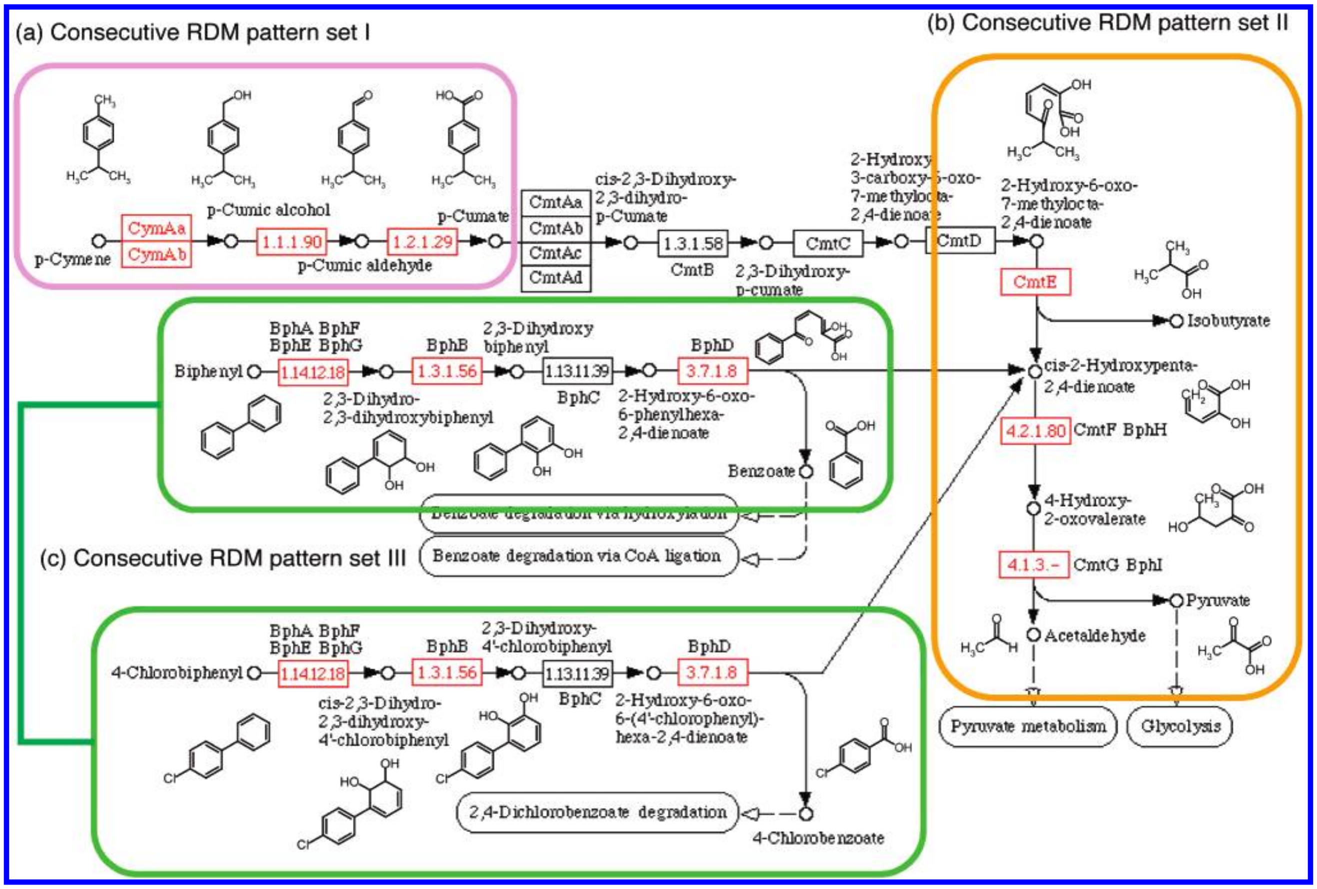

(a) Consecutive RDM pattern set I
(b) Consecutive RDM pattern set II
(c) Consecutive RDM pattern set III
p-Cymene
CymAa
CymAb
p-Cumic alcohol
1.1.1.90
p-Cumic aldehyde
1.2.1.29
p-Cumate
CmtAa
CmtAb
CmtAc
CmtAd
cis-2,3-Dihydroxy-2,3-dihydro-p-Cumate
1.3.1.58
CmtB
2,3-Dihydroxy-p-cumate
CmtC
2-Hydroxy-3-carboxy-6-oxo-7-methylocta-2,4-dienoate
CmtD
2-Hydroxy-6-oxo-7-methylocta-2,4-dienoate
CmtE
Isobutyrate
Biphenyl
BphA BphF
BphE BphG
1.14.12.18
BphB
1.3.1.56
2,3-Dihydro-2,3-dihydroxybiphenyl
2,3-Dihydroxy biphenyl
1.13.11.39
BphC
BphD
3.7.1.8
2-Hydroxy-6-oxo-6-phenylhexa-2,4-dienoate
Benzoate
Benzoate degradation via hydroxylation
Benzoate degradation via CoA ligation
cis-2-Hydroxypenta-2,4-dienoate
4.2.1.80
CmtF BphH
4-Hydroxy-2-oxovalerate
4.1.3.-
CmtG BphI
Pyruvate
Acetaldehyde
Pyruvate metabolism
Glycolysis
4-Chlorobiphenyl
BphA BphF
BphE BphG
1.14.12.18
BphB
1.3.1.56
cis-2,3-Dihydro-2,3-dihydroxy-4'-chlorobiphenyl
2,3-Dihydroxy-4'-chlorobiphenyl
1.13.11.39
BphC
BphD
3.7.1.8
2-Hydroxy-6-oxo-6-(4'-chlorophenyl)-hexa-2,4-dienoate
2,4-Dichlorobenzoate degradation
4-Chlorobenzoate

Figure 7. An example of consecutive distributions of characteristic RDM patterns in the KEGG biphenyl degradation pathway (map00621). The reactions with characteristic RDM patterns frequently occurring in this xenobiotics biodegradation are highlighted with colored boxes. The consecutive RDM pattern sets I, II, and III are represented by oblong boxes colored with purple, orange, and green, respectively. The consecutive RDM pattern set I is also observed in toluene and xylene degradation (map00622), 2,4-dichlorobenzoate degradation (map00623), and 1-and 2-methylnaphthalene degradation (map00624) pathways. The consecutive RDM pattern set II is also found in toluene and xylene degradation (map00622), styrene degradation (map00643), 1,4-dichlorobenzene degradation (map00627), ethylbenzene degradation (map00642), fluorene degradation (map00628), carbazole degradation (map00629), and benzoate via hydroxylation (map00362) pathways. The consecutive RDM pattern set III is also found in styrene degradation (map00643) and ethylbenzene degradation (map00642) pathways.

attempt to organize such knowledge based on the well-curated databases in KEGG for chemical compounds, reactions, and reactant pairs.

Genomic and Chemical Annotations. This process of organizing chemical knowledge can be compared with our other efforts to organize genomic knowledge in the process of computerized genome annotation, in which gene functions are assigned for completely sequenced genomes. In KEGG[3] the genome annotation is performed using ortholog annotation, where an ortholog is a group of functionally identical genes in different organisms. Genes in a newly sequenced genome are assigned KEGG Orthology (KO) identifiers or K numbers based on a computational method utilizing the well-curated databases, KEGG GENES and KO. Thus, the huge, chaotic space of all genes in all organisms is reduced to a more concise, well-organized space of orthologs, enabling better understanding of genomic information.

RDM patterns are similar, at least in spirit, to KO groups in which the space of all enzyme-catalyzed reactions is reduced to a more concise space of RDM patterns. RDM patterns are, however, less refined than KO groups at the moment. As shown in Table 1, the majority of RDM patterns correspond to single types of M atom changes; RD patterns may be sufficient in those cases. This also suggests the need for a similarity measure among RDM patterns. With further development of curated databases and refined computational methods, chemical annotation of metabolic compounds will become feasible.

As mentioned in the Introduction, there is another class of compounds, which we call regulatory compounds. They include exogenous compounds, such as drugs and environmental compounds. Some compounds may be both metabolic and regulatory in a single organism or in different organisms. For example, environmental compounds may be metabolized in bacteria but may be toxic in mammals. Because of a wide variety of interactions involved between these compounds and the biological macromolecules, different approaches will have to be taken for chemical annotation of regulatory compounds. We are currently accumulating data on known protein-small molecule interactions, such as enzyme inhibitors and receptor agonists/antagonists, in the KEGG BRITE database (http://www.genome.jp/kegg/brite.html). We also plan to analyze large-scale chemical genomics screening data to extract any empirical patterns that characterize regulatory compounds.

Improvements of the Prediction Scheme with Genomic Constraints. The space of enzyme-catalyzed reactions is a subset of all known organic reactions, and it is determined by the space of enzyme genes in all organisms. In the prediction scheme presented in this paper, we considered only the bacterial reactions (by removing mammalian cytochrome P450 reactions) when assigning plausible RDM patterns. Generally speaking, because whether reactions are plausible depends on the presence of enzyme genes, one possibility to improve the accuracy of our prediction system is to distinguish smaller organism groups. However, it is better to consider the bacterial community as a whole rather than individual species when examining xenobiotics degradation capability, because of the high rates of mutations and gene transfers.

In Figure 6 we investigated the distribution of RDM patterns in the three organism groups: eukaryotes, bacteria, and archaea. The pie chart indicates that specific RDM patterns are more abundant in bacteria than in eukaryotes or archaea. The pathway category of xenobiotics biodegradation contained many bacteria specific RDM patterns, such as for hydration of aromatic rings by means of dioxygenase (see Results).

We are currently extending our prediction scheme for use in the biosynthetic pathways of plant secondary metabolites. A number of phytochemical compound structures are already available in the KEGG COMPOUND database, but not much is known about the biosynthetic pathways and the responsible enzyme genes. It is obvious that the RDM pattern library not only needs to be extended to include additional reactions but also needs proper restriction based on the constraint of the genomic content of enzyme genes.

Improvements of the Prediction Scheme with Pathway Constraints. Another possibility to improve the accuracy of chemical annotation is to incorporate additional knowledge obtained from pathway data. There are cases where sets of consecutive RDM patterns are conserved in different pathways. Figure 7 shows such an example in the KEGG biphenyl degradation pathway in the category of xenobiotics biodegradation. Consecutive RDM pattern sets, designated by I, II, and III, were observed in this pathway, and each set was found to be conserved across different pathways (see legend). The consecutive pattern set I is an oxidation process where the alkyl group is oxidized to alcohol and carboxylic acid. The consecutive pattern set II yields acetaldehyde (or propanol) and pyruvate. The consecutive pattern set III is a dioxygenation process (using oxygen), which removes the aromaticity of aromatic rings and recovers the aromaticity. These consecutive pattern sets are associated with similar sets of consecutive EC numbers, and the corresponding genes for some of the consecutive pattern set II form operon structures. A comprehensive survey of the occurrences of such consecutive pattern sets is under way. In conclusion, an integrated analysis of chemical, genomic, and pathway information will enable extraction and understanding of biological information encoded in the chemical structures of small molecules.

ACKNOWLEDGMENT

We thank Rumiko Yamamoto, Tomoko Komeno, Junko Yabuzaki, and Yugo Shimizu for the curation of the KEGG RPAIR database and Nelson Hayes for critical reading of the manuscript. This work was supported by the 21st Century COE Program “Genome Science” and a grant-in-aid for scientific research on the priority area “Comprehensive Genomics”, both from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency. The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Supporting Information Available: Additional information for unique and frequently observed RDM patterns in the KEGG pathway categories. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Stockwell, B. R. Chemical genetics: ligand-based discovery of gene function. *Nat. Rev. Genet.* **2000**, *1*, 116−125.

(2) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824−828.

(3) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354−357.

(4) Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* **2003**, *125*, 11853−11865.

(5) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* **2004**, *126*, 16487−16498.

(6) Nobeli, I.; Thornton, J. M. A bioinformatician’s view of the metabolome. *BioEssays* **2006**, *28*, 534−545.

(7) McShan, D. C.; Updadhayaya, M.; Shah, I. Symbolic inference of xenobiotics metabolism. *Pac. Symp. Biocomput.* **2004**, *9*, 545−556.

(8) Langowski, J.; Long, A. Computer systems for the prediction of xenobiotic metabolism. *Adv. Drug Delivery Rev.* **2002**, *54*, 407−415.

(9) Klopman, G.; Tu, M. Structure-biodegradability study and computer-automated prediction of aerobic biodegradation of chemicals. *Environ. Toxicol. Chem.* **1997**, *16*, 1829−1835.

(10) Talafous, J.; Sayre, L. M.; Mieyal, J. J.; Klopman, G. META.2. A dictionary model of mammalian xenobiotic metabolism. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1326−1333.

(11) Darvas, F. Predicting metabolic pathways by logic programming. *J. Mol. Graphics Modell*. **1988**, *6*, 80−86.

(12) Ellis, L. B. M.; Roe, D.; Wackett, L. P. The university of minnesota biocatalysis/biodegradation database: the first decade. *Nucleic Acids Res.* **2006**, *34*, D517−D521.

(13) Hou, B. K.; Ellis, L. B. M.; Wackett, L. P. Encoding microbial metabolic logic: predicting biodegradation. *J. Ind. Microbiol. Biotechnol.* **2004**, *31*, 261−272.

(14) Barrett, A. J.; Canter, C. R.; Liebecq, C.; Moss, G. P.; Saenger, W.; Sharon, N.; Tipton, K. F.; Vnetianer, P.; Vliegenthart, V. F. G. *Enzyme Nomenclature*; Academic Press: San Diego, California, 1992.

(15) Caraux, G.; Pinloche, S. PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics* **2005**, *21*, 1280−1281.

(16) Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14863−14868.

(17) Potrawfke, T.; Timmis, K. N.; Wittich, R. M. Degradation of 1,2,3,4-tetrachlorobenzene by pseudomonas chlororaphis RW71. *Appl. Environ. Microbiol.* **1998**, *64*, 3798−3806.

(18) Mayeno, A. N.; Yang, R. S. H.; Reisfeld, B. Biochemical reactions network modeling: predicting metabolism of organic chemical mixtures. *Environ. Sci. Technol.* **2005**, *39*, 5363−5371.

(19) Janssen, D. B.; Dinkla, I. J. T.; Poelarends, G. J.; Terpstra, P. Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environ Microbiol.* **2005**, *7*, 1868−1882.

(20) Werlen, C.; Kohler, H. P.; van der Meer, J. R. The broad substrate chlorobenzene dioxygenase and cis-chlorobenzene dihydrodiol dehydrogenase of *Pseudomonas* sp. strain P51 are linked evolutionarily to the enzymes for benzene and toluene degradation. *J. Biol. Chem.* **1996**, *271*, 4009−4016.

(21) van der Meer, J. R.; van Neerven, A. R.; de Vries, E. J.; de Vos, W. M.; Zehnder, A. J. Identification of a novel composite transposable element, Tn5280, carrying chlorobenzene dioxygenase genes of *Pseudomonas* sp. strain P51. *J. Bacteriol.* **1991**, *173*, 6−15.

(22) Middeldorp, P. J. M.; Jaspers, M.; Zehnder, A. J. B.; Schraa, G. Biotransformation of α-, *β*-, *γ*- and *δ*-Hexachlorocyclohexane under methanogenic conditions. *Environ. Sci. Technol.* **1996**, *30*, 2345−2349.

(23) Raschke, H.; Fleischmann, T.; van der Meer, J. R.; Kohler, H. P. *cis*-Chlorobenzene dihydrodiol dehydrogenase (TcbB) from *Pseudomonas*

sp. strain P51, expressed in *Escherichia coli* DH5alpha(pTCB149), catalyzes enantioselective dehydrogenase reactions. *Appl. Environ. Microbiol.* **1999**, *65*, 5242−5246.

(24) Potrawfke, T.; Armengaud, J.; Wittich, R. M. Chlorocatechols substituted at positions 4 and 5 are substrates of the broad-spectrum chlorocatechol 1,2-dioxygenase of Pseudomonas chlororaphis RW71. *J. Bacteriol.* **2001**, *183*, 997−1011.

(25) Hammer, A.; Hildenbrand, T.; Hoier, H.; Ngai, K. L.; Schlomann, M.; Stezowski, J. J. Crystallization and preliminary X-ray data of chloromuconate cycloisomerase from Alcaligenes eutrophus JMP134 (pJP4). *J. Mol. Biol.* **1993**, *232*, 305−307.

(26) Ngai, K. L.; Schlomann, M.; Knackmuss, H. J.; Ornston, L. N. Dienelactone hydrolase from *Pseudomonas* sp. strain B13. *J. Bacteriol.* **1987**, *169*, 699−703.

(27) Tompkins, F. C., Jr.; Goldsmith, R. L. A new personal dosimeter for the monitoring of industrial pollutants. *Am. Ind. Hyg. Assoc. J.* **1977**, *38*, 371−377.

(28) Hashimoto, K.; Goto, S.; Kawano, S.; Aoki-Kinoshita, K. F.; Ueda, N.; Hamajima, M.; Kawasaki, T.; Kanehisa, M. KEGG as a glycome informatics resource. *Glycobiology* **2006**, *16*, 63R−70R.

(29) Kawano, S.; Hashimoto, K.; Miyama, T.; Goto, S.; Kanehisa, M. Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics* **2005**, *21*, 3976−3982.

(30) Minowa, Y.; Araki, M.; Kanehisa, M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.* **2007**, in press.

CI700006F