JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

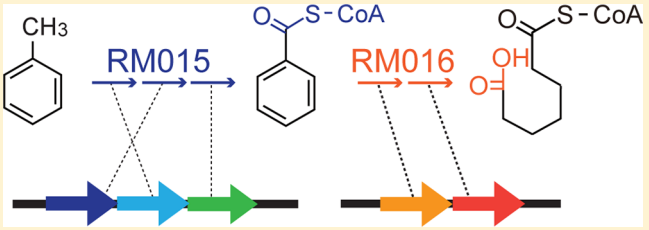
Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions

Ai Muto, Masaaki Kotera, Toshiaki Tokimatsu, Zenichi Nakagawa, Susumu Goto, and Minoru Kanehisa*

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

Supporting Information

ABSTRACT: The metabolic network is both a network of chemical reactions and a network of enzymes that catalyze reactions. Toward better understanding of this duality in the evolution of the metabolic network, we developed a method to extract conserved sequences of reactions called reaction modules from the analysis of chemical compound structure transformation patterns in all known metabolic pathways stored in the KEGG PATHWAY database. The extracted reaction modules are repeatedly used as if they are building blocks of the metabolic network and contain chemical logic of organic reactions. Furthermore, the reaction modules often correspond to traditional pathway modules defined as sets of enzymes in the KEGG MODULE database and sometimes to operon-like gene clusters in prokaryotic genomes. We identified well-conserved, possibly ancient, reaction modules involving 2-oxocarboxylic acids. The chain extension module that appears as the tricarboxylic acid (TCA) reaction sequence in the TCA cycle is now shown to be used in other pathways together with different types of modification modules. We also identified reaction modules and their connection patterns for aromatic ring cleavages in microbial biodegradation pathways, which are most characteristic in terms of both distinct reaction sequences and distinct gene clusters. The modular architecture of biodegradation modules will have a potential for predicting degradation pathways of xenobiotic compounds. The collection of these and many other reaction modules is made available as part of the KEGG database.



INTRODUCTION

Metabolism is the most basic aspect of life. It represents a chemical system generating all necessary chemical substances in living cells through chemical reactions. It also represents a genetic system in the sense that chemical reactions are catalyzed by genome-encoded enzymes. The dual aspect of metabolism has been utilized for metabolic reconstruction, where the repertoire of enzyme genes in the genome is used to infer chemical capacity of an organism, such as biosynthetic and biodegradation potentials and environmental adaptability. The procedure for metabolic reconstruction generally involves finding orthologous genes in the genome that match the network of enzymes in known metabolic pathways.[1−4] This type of analysis has revealed conserved functional units in metabolic pathways; for example, sets of enzyme genes adjacent on the chromosome encoded in operon-like structures were linked to functional units of successive reaction steps.[5] In this paper we focus on the chemistry of metabolic reactions. Our working hypothesis is that the functional units revealed by enzyme clusters must also reflect chemical units of organic reactions. We have thus developed a method to extract, what we call, reaction modules without using the data on enzymes and enzyme genes.

The metabolic pathway reconstruction problem is a special case of the pathway alignment problem, where the pathway similarity is defined by the sequence similarity of orthologous enzyme genes.[3] In another example of pathway alignment, Tohsato et al.[6] used the EC number (Enzyme Commission number) similarity to find similar reaction steps in the well-characterized metabolic network. Pinter et al.[7] followed Tohsato's definition of EC number similarity, and developed methods to detect conserved metabolic pathways among different organisms and divergent pathways within an organism. Wernicke et al.[8] followed Pinter's definition of aligning metabolic pathways and provided a faster algorithm. In yet another example of pathway alignment, Tohsato and Nishimura[9] used the similarity of substructure changes to detect similar sequences of metabolites along known metabolic pathways. Ay et al.[10] developed a method that combined EC number similarity and metabolite similarity, as well as the metabolic network topology similarity to conduct the pathway alignment.

Not surprisingly, the outputs of these different pathway alignment studies are different. The assignment of orthologous enzyme genes can only deal with the pathways that consist of the same reactions catalyzed by the same enzymes. The use of EC number similarity allows not only the same reactions but also somewhat different reactions to be considered because of the EC number hierarchy. However, since the EC numbers are manually given to experimentally characterized enzymes with varying standards, the EC number similarity is not a reliable

Received: November 9, 2012
Published: February 5, 2013

dx.doi.org/10.1021/ci3005379 | J. Chem. Inf. Model. 2013, 53, 613−622

measure for systematic analysis, for example, comparison of genomic diversity of enzyme genes and chemical diversity of enzyme-catalyzed reactions. Furthermore, since the EC numbers are annotated to genes in a genome according to sequence similarity, their use is not much different from ortholog-based approaches in many types of analyses.

Here we introduce a new similarity measure for pathway alignment. It is based on the similarity of chemical structure transformation patterns along the metabolic pathways. This is a purely chemical similarity measure without incorporating any protein sequence information or the EC number information, enabling the analysis of reactions with no EC numbers assigned or even with no enzymes identified. The new measure applies to localized structural changes extracted from compound pairs (substrate−product pairs) accommodating global structure differences of individual compounds. Our data set is derived from the KEGG database,[11] more specifically the KEGG RPAIR (reactant pair) database[12] and the KEGG RCLASS (reaction class) database. KEGG RPAIR contains chemical structure transformation patterns in all known enzyme-catalyzed reactions, and KEGG RCLASS consists of reaction class (RC) entries defined by the identity of chemical transformation patterns in the "main" reactant pairs that generally correspond to the main substrate and product pairs shown on the KEGG metabolic pathway maps. The reaction modules, which are conserved sequences of similar reactions, are systematically searched in the KEGG metabolic pathways using the similarity scoring scheme between reaction class entries. Extracted reaction modules are then compared with the pathway modules in the KEGG MODULE database, which are defined as sets of enzyme orthologs represented by the KEGG Orthology (KO) entries.[11]

MATERIALS AND METHODS

Metabolic Pathway Database. The present analysis is based on the KEGG database (http://www.kegg.jp/) release 62.0+ (May 24, 2012). We used the metabolic pathways stored in the Metabolism section of the KEGG PATHWAY database, a collection of manually drawn pathway maps representing knowledge on the molecular interaction and reaction networks summarized from literature. Generally, KEGG pathway maps are not organism specific and contain integrated information from different organisms. The KEGG "reference" pathway maps are drawn in a generic form in terms of orthologs, so that they can be expanded to organism-specific pathways by combining with the gene information in individual genomes. A typical KEGG metabolic pathway map depicts how small molecules (drawn as circles) are converted by reactions catalyzed by enzymes (drawn as rectangles). The dual aspect of metabolism is annotated in KEGG by linking each rectangle in the KEGG reference metabolic pathway map to both the enzyme ortholog entry (identified by K number) in the KEGG Orthology (KO) database and the reaction entry (identified by R number) in the KEGG REACTION database. The KEGG pathway maps in an XML format (KGML files) contain information about the relationship of neighboring rectangles, which has been utilized to extract consecutive reaction steps as sequences of R numbers.

Reactant Pairs. The KEGG REACTION database contains all known enzymatic reactions taken from the Enzyme Nomenclature[13] and also from the metabolic pathway section of the KEGG PATHWAY database. The KEGG release that we used contains 8990 reactions including 4321 enzyme nomenclature reactions. Among them, 6238 reactions including 2595 enzyme nomenclature reactions appear on the KEGG metabolic pathways. Generally, one reaction consists of multiple substrates and products. Reactant pairs are defined as one-to-one relationships of substrate−product pairs by considering the flow of atoms (other than hydrogen atoms) in enzymatic reactions as well as the six EC number classes.[12] There were 13 448 reactant pairs stored in the KEGG RPAIR database.

Each reactant pair is associated with the chemical transformation pattern in the RDM notation consisting of the KEGG atom type changes at the reaction center (R), the difference substructure (D), and the matching substructure (M) atoms (see the Supporting Information for more details). The RDM notation for the pair of reactants A and B is as follows:

$$\mathrm{RDM(A, B)} = \mathrm{RA{-}RB{:}\ DA{-}DB{:}\ MA{-}MB}$$

For example, a typical acyltransferase reaction on primary amine is described as

$$\mathrm{RDM} = \mathrm{N1a{-}N1b{:}\ *{-}C5a{:}\ C1b{-}C1b}$$

The KEGG atom type[14] generally consists of three characters, such as N1a for primary amine nitrogen and C5a for ketone carbon. The first (upper case letter) indicates the atomic species, the second (numeral) represents the predefined class of atomic bonding for each atomic species, and the third (lower case letter) represents the predefined class of topological information, e.g., the number of substituted groups. The total of 68 atom types have been defined to distinguish important functional groups in biological small molecules. In the above notation, * corresponds to a hydrogen atom (not defined in the KEGG atom types) for substitution reaction. Note that * sometimes means that there is no atom in the D-substructure. Note also that the RDM notation generally represents a single chemical bond that changes during a reaction; in the case where a reaction generates or degrades more than one chemical bond, then more than one RDM notation is required.[12]

Reaction Class. The KEGG RCLASS database has been developed to classify chemical structure transformation patterns associated with all the reactions that appear in the KEGG metabolic pathway maps. The database is a collection of reaction class entries (identified by RC numbers), each representing a unique RDM chemical structure transformation pattern for a group of "main" reactant pairs in the KEGG RPAIR database. In other words, each RCLASS entry indicates a collection of substrate−product pairs that appear on the KEGG metabolic pathway maps and whose chemical transformation patterns are identical. The RCLASS entry is computationally generated from the KEGG RPAIR database and manually annotated with a diagram of chemical transformation pattern and other information. There were 2481 RCLASS entries in this study.

RESULTS

Similarity Grouping of RCLASS Entries. Because the RDM chemical transformation patterns and the resulting RCLASS entries are too finely classified, we first introduced a similarity scoring scheme for RCLASS entries in order to detect similar (in addition to identical) chemical transformation patterns. This is based on the fingerprint representation of KEGG atom types using twelve keys. The keys indicate the presence or absence of a carbon atom, a carbon atom having a π

dx.doi.org/10.1021/ci3005379 | J. Chem. Inf. Model. 2013, 53, 613−622

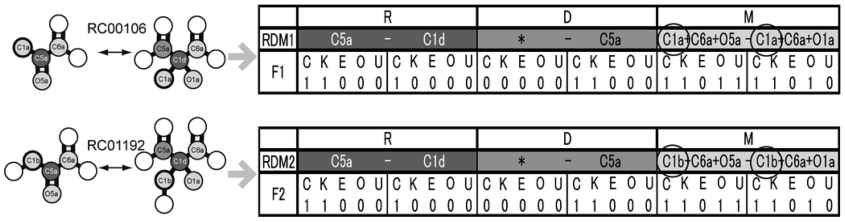


Figure 1. Fingerprint representation of the RDM pattern. Two reaction class entries RC00106 and RC01192 are shown for the reactant pairs of pyruvate and acetolactate (upper) and oxobutanoate and 2-aceto-2-hydroxybutanoate (lower). The RDM notation for these reaction class entries is converted to the fingerprint representation, which reveals that they are the same in the fingerprint representation despite the difference in the M atom (circled) in the RDM notation.

bond, a carbon atom in a carbonyl group, an oxygen atom, an oxygen atom with unpaired electron, a nitrogen atom, a phosphorus atom, a sulfur atom, a halogen atom, other atoms, an atom in aromatic ring, and an atom in any ring (see the Supporting Information for more details). Figure 1 illustrates how the RDM notation is converted into the 72-bit fingerprint notation. For example, methyl (C1a), methylene (C1b), and other sp3 carbon atoms (C1c and C1d) are given different KEGG atom types, but they are the same in the fingerprint representation.

The similarity score S between two RCLASS entries, each consisting of a single RDM pattern, is defined as

$$S(\mathrm{RC}_1, \mathrm{RC}_2) = w_\mathrm{R} J_a(v_{\mathrm{R}1}, v_{\mathrm{R}2}) + w_\mathrm{D} J_a(v_{\mathrm{D}1}, v_{\mathrm{D}2}) + w_\mathrm{M} J_a(v_{\mathrm{M}1}, v_{\mathrm{M}2})$$

where the fingerprint v is compared separately for the R, D, and M atoms, and the average Jaccard's coefficient Ja is weighted depending on, for example, whether the D atom is missing (see the Supporting Information for more details). In the present analysis, we use the similarity threshold score of 1.0; namely, we simply use the criterion of the same fingerprint to group RCLASS entries into a similarity group. As a result, 2481 RCLASS entries were grouped into 376 similarity groups with more than one member and the remaining 1190 singletons.

Extraction of Conserved RCLASS Sequence Patterns. On the basis of the similarity measure among RCLASS entries as defined above, we extracted "reaction modules", which in our definition are consecutive reaction steps (reaction sequences) with conserved RCLASS sequence patterns that are observed in different metabolic pathways. We used the following procedure to extract such conserved patterns (see the Supporting Information for more details). Known metabolic pathways in the KEGG PATHWAY database are split into all possible subsequences of 2−8 consecutive reactions. The pathways involving branches are split into all combinations of linear reaction sequences. The direction of the reaction pathway is taken into consideration when generating consecutive reaction sequences that contain irreversible reactions. When the pathway forms a cycle, all possible reaction sequences that do not include the same metabolite more than once are generated.

For a given length between 2 and 8, the reaction (R number) sequences thus generated were first converted to the RCLASS (RC number) sequences. Two RCLASS sequences are considered to be identical when the corresponding RC numbers are the same, and to be similar when they belong to the same similarity group in the fingerprint representation. For each given length, conserved RCLASS sequence patterns consisting of such similarity groups were extracted from the entire collection of KEGG metabolic pathways. The result is shown in Table 1 indicating roughly one-half of the pathways correspond to conserved reaction sequence patterns. After computationally removing shorter patterns embedded in longer patterns, we manually examined the results to identify reaction modules.

Table 1. Number of Conserved RCLASS Sequence Patterns Found in the KEGG Metabolic Pathways

length	no. of conserved patterns	no. of reactions included	coverage[a]
2	928	3479	0.599
3	770	2503	0.431
4	534	1662	0.286
5	338	1074	0.185
6	218	765	0.132
7	140	527	0.091
8	88	399	0.069
total	3016		

[a]The ratio to 5805 reactions, the total number of reactions with RC assignment in the KEGG pathways.

General Characteristics of Reaction Modules. The list of manually refined reaction modules is partially shown in Table 2 and fully shown at http://www.kegg.jp/kegg/reaction/rmodule.html. We found three general characteristics of reaction modules. First, reaction modules are repeatedly used in different pathways to generate different chemical substances. Second, reaction modules are used in combination as if they are building blocks of the metabolic network. Third, and most importantly, reaction modules (also called RC modules) derived from chemical properties of substrate−product structure transformation patterns tend to correspond to KEGG pathway modules (also called KO modules) defined as sets of enzyme orthologs in the genome, especially gene clusters in operon-like structures coding for the enzymes. The total of 26 corresponding KO modules were found for 16 out 21 RC modules shown in Table 2, and all the KO modules except one contained operon-like gene clusters in some genomes (for an updated list, see http://www.kegg.jp/brite/ko00003). Thus, we confirmed our working hypothesis mentioned in the Introduction; the functional units revealed by enzyme clusters reflect chemical units of organic reactions. Here we report detailed analysis of the reaction modules for 2-oxocarboxylic acid chain extension and modification, fatty acid synthesis and beta-oxidation, and aromatic ring cleavage.

2-Oxocarboxylic Acid Chain Extension. One of the most characteristic reaction modules was RM001 for the chain extension of 2-oxocarboxylic acids, an important class of precursor metabolites. This module corresponds to the well-known sequence of reactions involving citrate and other

dx.doi.org/10.1021/ci3005379 | J. Chem. Inf. Model. 2013, 53, 613−622

Table 2. List of Reaction Modules Discussed in This Paper[a]

RC module	description	length
RM001	2-oxocarboxylic acid chain extension by tricarboxylic acid pathway	5
RM002	caboxyl to amino conversion using protective *N*-acetyl group	5
RM032	caboxyl to amino conversion	3
RM033	branched chain addition	4
RM030	glucosinolate biosynthesis	5
RM021	fatty acid synthesis using malonyl-CoA	4
RM020	fatty acid synthesis using acetyl-CoA (reversal of beta oxidation)	4
RM018	beta oxidation in acyl-CoA degradation	4
RM003	methyl to carboxyl conversion on aromatic ring, aerobic	3
RM004	dihydroxylation of aromatic ring, type 1 (dioxygenase and dehydrogenase reactions)	2
RM005	dihydroxylation of aromatic ring, type 1a (dioxygenase and decarboxylating dehydrogenase reactions)	2
RM006	dihydroxylation of aromatic ring, type 2 (two monooxygenase reactions)	2
RM008	ortho-cleavage of catechol (beta-ketoadipate pathway)	4
RM009	meta-cleavage of catechol	6
RM010	dihydroxylation and meta-cleavage of aromatic ring, type 1	4
RM015	oxidation of methyl group on aromatic ring, anaerobic	6
RM016	aromatic ring cleavage via beta oxidation, anaerobic	3
RM025	conversion of amino acid moiety to carboxyl group	3
RM022	nucleotide sugar biosynthesis, type 1	3
RM023	nucleotide sugar biosynthesis, type 2	2
RM027	hydroxylation and methylation motif	2

[a]See http://www.kegg.jp/kegg/reaction/rmodule.html for the full list of reaction modules.

tricarboxylic acids in the TCA cycle (map00020 in KEGG), where an acetyl-CoA derived carbon is used to extend the 2-oxocarboxylic acid chain from oxaloacetate (2-oxobutanedioate) to 2-oxoglutarate, namely, from a four-carbon (C4) compound to a five-carbon (C5) compound. This is in fact the only part in the TCA cycle that involves tricarboxylic acids. Interestingly, we identified three more examples of the same reaction module RM001. One is a further extension from 2-oxoglutarate (C5) to 2-oxoadipate (C6) in lysine biosynthesis pathway (map00300). Another is found in valine, leucine, and isoleucine biosynthesis pathway (map00290) where pyruvate (2-oxopropanoate) is extended to 2-oxobutanoate, and 2-oxoisovalerate is extended to 2-oxoisocaproate.[15] Furthermore, in the biosynthesis pathway of glucosinolates (map00966), which are plant secondary metabolites, a six tandem repeat of RM001 is found from 2-oxo-4-methylthiobutanoate to 2-oxo-10-methylthiodecanoate (see Supporting Information Figure S1).

These additional examples were found by considering not only the similarity grouping of reaction class entries, but also the matching of multistep reactions to an overall reaction. In the KEGG pathway map for the citrate cycle (map00020), the conversion from oxaloacetate to 2-oxoglutarate (RM001) is shown as follows: oxaloacetate and acetyl-CoA generating citrate (RC00067), converting to cis-aconitate (RC00498), converting to isocitrate (RC00618), and converting to 2-oxoglutarate in two reaction steps (RC00084+RC00626) or in one step (RC00114). The variation of the last conversion originates from the definition of EC numbers: EC 1.1.1.42 defined as two step reactions and EC 1.1.1.41 defined as one overall reaction (RC00114). Furthermore, the second and third reactions (RC00498+RC00618) are catalyzed by the same enzyme EC 4.2.1.3, although the overall one-step reaction from citrate to isocitrate (RC00497) is not included in the KEGG map. As shown in Table 3, the reaction module RM001 apparently consists of different RCLASS sequences, but they are actually the same with these considerations.

Modification of 2-Oxocarboxylic Acids. The reaction module RM001 for 2-oxocarboxylic acid chain extension by tricarboxylic acid pathway was found to be used in combination with three modification modules, RM002 (including RM032), RM033, and RM030, together with a reductive amination step (RC00006 or RC00036). In Figure 2, RM002 is for conversion of carboxyl group to amino group in the biosynthesis of basic amino acids (ornithine and lysine), and RM033 is for addition of branched chains in the biosynthesis of branched-chain amino acids (valine, leucine, and isoleucine). RM030 in glucosinolate biosynthesis pathway is for conversion to oxime followed by addition of thio-glucose moiety (see Supporting Information Figure S2).

For the chemical modification module RM002 in Table 3 and Figure 2, the reaction sequence contains addition and

Table 3. Reaction Modules Involving 2-Oxocarboxylic Acids

RC module	pathway	overall reaction	RCLASS sequence
RM001	citrate cycle (map00020)	oxaloacetate → 2-oxoglutarate	RC00067 RC00498 RC00618 RC00084+RC00626
	lysine biosynthesis (map00300)	2-oxoglutarate → 2-oxoadipate	RC00067 RC00498 RC00618 RC00114
	isoleucine biosynthesis (map00290)	pyruvate → 2-oxobutanoate	RC01205 RC00976 RC00977 RC00417
	leucine biosynthesis (map00290)	2-oxoisovalerate → 2-oxoisocaproate	RC00470 RC01041 RC01046 RC00084+RC00577
	glucosinolate biosynthesis (map00966)	2-oxo-4-methylthiobutanoate → 2-oxo-10-methylthiodecanoate	RC00067 RC00497 RC00114 (six repeats)
RM002	lysine biosynthesis (map00300)	2-aminoadipate → lysine	RC00064 RC00043 RC00684 RC00062 RC00064
	arginine biosynthesis (map00330)	glutamate → ornithine	RC00064 RC00043 RC00684 RC00062 RC00064
RM032	ectoine biosynthesis (map00260)	aspartate → 2,4-diaminobutanoate	RC00043 RC00684 RC00062
RM033	valine biosynthesis (map00290)	pyruvate → 2-oxoisovalerate	RC01192 RC00837 RC00726 RC00468
	isoleucine biosynthesis (map00290)	2-oxobutanoate → 3-methyl-2-oxopentanoate	RC01192 RC01726 RC00726 RC01714
RM030	glucosinolate biosynthesis (map00966)	homomethionine → glucoiberverin	RC02295 RC02210 RC02265 RC00882 RC00883

dx.doi.org/10.1021/ci3005379 | J. Chem. Inf. Model. 2013, 53, 613−622

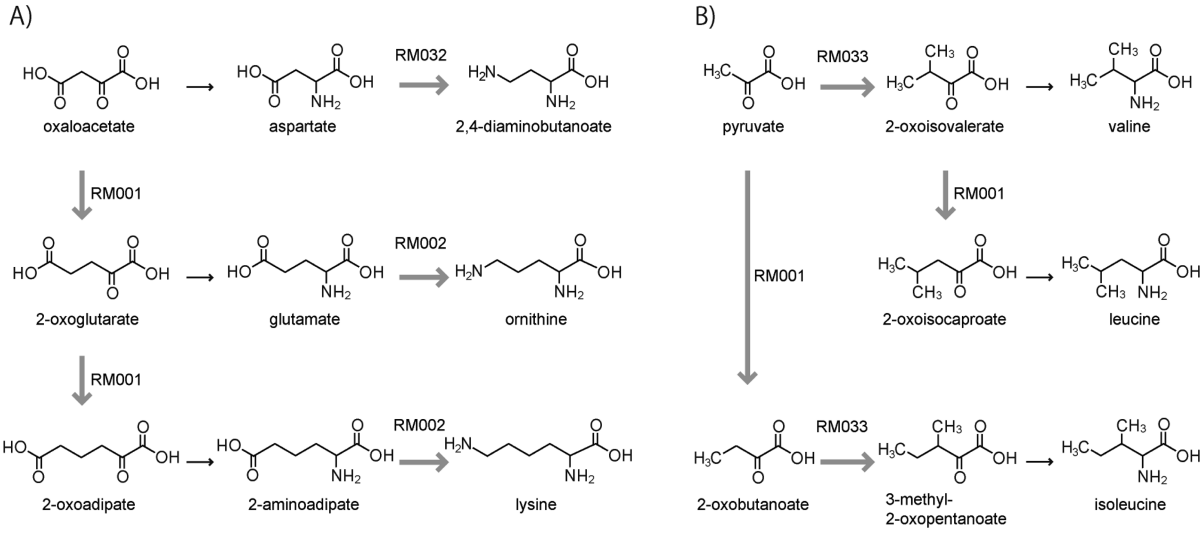


Figure 2. Architecture of reaction modules consisting of 2-oxocarboxylic acid chain extension and modification, generating (A) basic amino acids and (B) branched-chain amino acids. Vertical arrows indicate the extension modules RM001. Horizontal arrows indicate the modification modules RM002, RM032, and RM033 together with the reductive amination step (RC00006 and RC00036).

removal of *N*-acetyl group as a protective group in the first and last steps (RC00064). We found the same overall reaction sequence without the protective group (RM032) in the conversion from aspartate to 4-diaminobutanoate in ectoine biosynthesis pathway (map00260) (see Figure 2 and Table 3). This seems to suggest an interesting variation, possibly an evolution, of reaction modules. While RM032 (RC00043 RC00684 RC00062) is used for shorter carboxylic acid chains, RM002 (RC00064 RC00043 RC00684 RC00062 RC00064), containing *N*-acetylation and *N*-deacetylation steps surrounding the core sequence of RM032, is used for longer chains. Furthermore, in the lysine biosynthesis pathway from 2-aminoadipate to lysine, the surrounding steps are proteinaceous *N*-terminal modifications.[16] Thus, basically the same overall reaction is achieved with varying degrees of complexity, namely the involvement of a small modified group and a modified group attached to a carrier protein, which may indicate a chemical evolution of the metabolic network to cope with longer chains (see Supporting Information Figure S3).

Reaction Modules Encoded in Enzyme Gene Clusters. The KEGG pathway modules (KO modules) in the KEGG MODULE database are represented by manually defined sets of enzyme orthologs, which often correspond to operon-like gene clusters in prokaryotic genomes. We examined relationships between the reaction modules (RC modules) extracted in the present analysis and the previously defined KO modules, especially the gene clusters that constitute the KO modules, to exclude the ambiguity of manual definition. We found, for example, that the RC module RM001 coincided well with the KO modules M00010, M00432, and M00535. As shown in Table 4, in the genome of *Pyrococcus furiosus*,[17] two gene clusters correspond to the reaction module RM001: the gene cluster (PF0203 PF0201 PF0202) for the RCLASS sequence (RC00067 RC00498+RC00618 RC00084+RC00626) in citrate cycle and the gene cluster (PF0937 PF0938+PF0939 PF0940) for the RCLASS sequence (RC00470 RC01041+RC01046 RC00084+RC00577) in leucine biosynthesis. It is interesting to note that although the first enzymes (citrate synthase PF0203 and 2-isopropylmalate synthase PF0937) do not share any sequence similarity, the second hydratases (PF0201 and PF0938+PF0939) and the third dehydrogenases (PF0202 and PF0940) form paralogous gene groups with sequence identity of around 35%, suggesting a link between genomic diversity and chemical diversity. Table 4 shows only a few examples of enzyme gene clusters for RM001 and RM002. Many more examples can be found in the KEGG database from the Ortholog table view of the KEGG MODULE entries (each entry is accessible at http://www.kegg.jp/module/M00010, etc.).

Table 4. Reaction Modules Corresponding to Enzyme Gene Clusters

RC module	overall reaction	KO module	gene cluster example[a]
RM001	oxaloacetate → 2-oxoglutarate	M00010	(pfu) PF0203 PF0201 PF0202
	2-oxoisovalerate → 2-oxoisocaproate	M00432	(pfu) PF0937 PF0938+PF0939 PF0940
	pyruvate → 2-oxobutanoate	M00535	(bth) BT_1858 BT_1860+BT_1859 BT_1857
RM002	2-aminoadipate → lysine	M00028	(bsu) BSU11200 BSU11210+BSU11190 BSU11220
	glutamate → ornithine	M00031	(ttr) Tter_0315+Tter_0316 Tter_0320 Tter_0319 Tter_0321 Tter_0317

[a]KEGG organism codes are shown in parentheses: pfu (T00075), *Pyrococcus furiosus* DSM 3638; bth (T00122), *Bacteroides thetaiotaomicrometer* VPI-5482; bsu (T00010), *Bacillus subtilis* 168; ttr (T01134), *Thermobaculum terrenum* ATCC BAA-798.

Fatty Acid Synthesis and Beta Oxidation. Fatty acids are long-chain carboxylic acids. It is well-known that fatty acid synthesis is a repetition of the four-step reaction sequence consisting of ketoacyl synthase (KS), ketoreductase (KR), dehydratase (DH), and enoylreductase (ER) reactions and each sequence increasing the acyl chain length by two. The four enzymes may come from four separate genes, two genes containing KS + KR and DH + ER domains or a single gene containing all domains. Thus, there seems to be an evolution of the genetic aspect of fatty acid synthesis. It is less known that there is a minor pathway for fatty acid biosynthesis. In contrast

dx.doi.org/10.1021/ci3005379 | J. Chem. Inf. Model. 2013, 53, 613−622

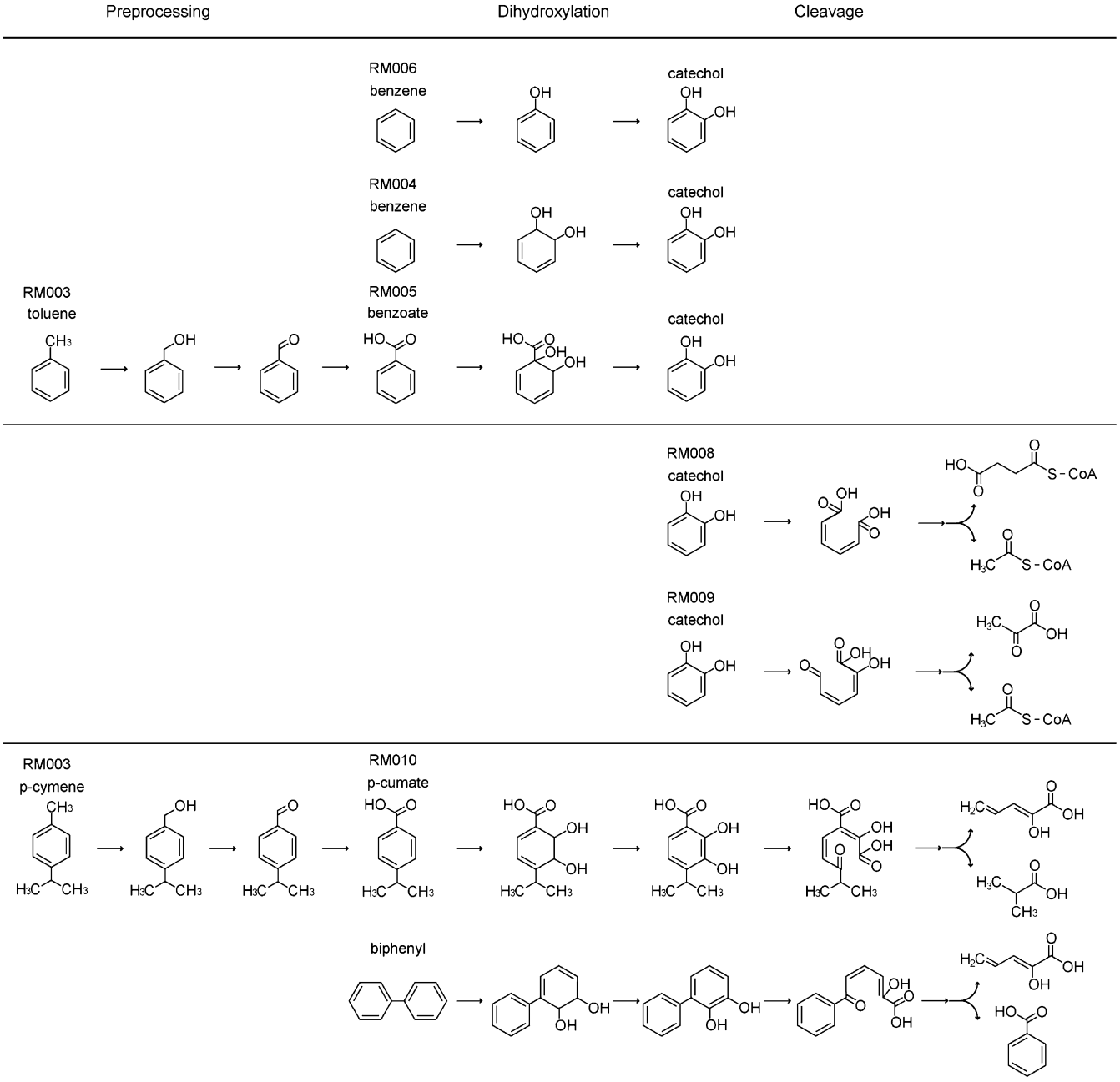


Figure 3. Aromatic ring cleavage modules in microbial degradation pathways. Aromatic rings are cleaved in the following three steps. The first step is an occasional preprocessing step (RM003) converting a methyl group into a carboxylic group on the aromatic ring. The second step is the main step of dihydroxylation, which is classified into three types (RM006, RM004, and RM005) depending on how two hydroxyl groups are added on the aromatic ring. The third step is either ortho-cleavage (RM008) or meta-cleavage (RM009) followed by characteristic reaction patterns leading to TCA cycle intermediates.

to the major pathway (map00061) that involves acyl carrier protein and uses malonyl-CoA as a carbon source, the minor pathway (map00062) in mitochondria does not involve acyl carrier protein and uses acetyl-CoA as a carbon source. The minor pathway is essentially the reversal of beta-oxidation (RM018) in fatty acid degradation (see below). The reaction modules RM021 and RM020, for the major and minor pathways respectively, consist of the following RCLASS sequences (see also Supporting Information Figure S4).

RM021: (RC02729 RC00117 RC00831 RC00076) for $n = 2$
(RC02728 RC00029 RC01095 RC00052) for $n \geq 4$

RM020: (RC00326 RC00117 RC00831 RC00076) for $n = 2$
(RC00326 RC00029 RC01095 RC00052) for $n \geq 4$

Except for the first KS reaction step distinguishing malonyl-[acp] and acetyl-CoA, the two modules RM021 and RM020 are essentially the same. We consider that the involvement of the acyl carrier protein is an evolution of the chemical aspect of fatty acid synthesis, possibly increasing the specificity and efficiency of reactions.

The beta oxidation module RM018 (the reversal of RM020) consists of the following RCLASS sequences:

dx.doi.org/10.1021/ci3005379 | J. Chem. Inf. Model. 2013, 53, 613−622

RM018: (RC00052 RC01095 RC00029 RC00326) for $n \geq 4$
(RC00076 RC00831 RC00117 RC00326) for $n = 2$

In addition to the fatty acid degradation pathway (map00071), RM018 is found in caprolactum degradation (map00930) from adipyl-CoA to succinyl-CoA and isoleucine degradation (map00280) from 2-methylbutanoyl-CoA to propanoyol-CoA (see Supporting Information Figure S5). A variation of RM018, with a somewhat different cleavage reaction, was found in the anaerobic benzoate degradation pathway (map00362) from pimeloyl-CoA to glutaryl-CoA and in primary bile acid synthesis (map00120) leading to chenodeoxycholoyl-CoA and choloyl-CoA.

More interestingly a distantly related module, RM016 (RC02034 RC00154 RC01429), is found in the anaerobic benzoate degradation pathway (map00362) and the limonene and pinene degradation pathway (map00903). RM016 is for the aromatic ring cleavage (hydrolysis) rather than the acyl chain cleavage (thiolysis), but the strategy to add oxygen preceding the cleavage reaction is similar. The RC similarity score was 0.8 for RC02034 and RC01095 (or RC00831) and 0.7 for RC00154 and RC00029 (or RC00117). Furthermore, the genes for RC00154 in the anaerobic benzoate degradation pathway and for RC00029 in the fatty acid biosynthesis pathway share sequence similarity.[18]

Aromatic Ring Cleavage in Microbial Biodegradation Pathways. Microorganisms are known to be capable of degrading diverse chemical substances including nonbiological chemicals in the environment that are mostly aromatic compounds. Figure 3 illustrates a representative set of reaction modules for biodegradation of aromatic compounds consisting of ring dihydroxylation modules (RM004−RM006), cleavage modules (RM008 and RM009), and a preprocessing module (RM003) for oxyfunctionalization, i.e., converting methyl group to carboxyl group on the aromatic ring. RM003 consists of a monooxygenase reaction (EC 1.14.13) and two steps of dehydrogenase reactions (EC 1.1.1 and 1.2.1) and is followed by a dihydroxylation module, as shown in Figure 3 for toluene to benzoate and *p*-cymene to *p*-cumate conversions.

On the basis of the similarity grouping of RCLASS entries, we categorized dihydroxylation reaction steps into three main modules (RM004, RM005, and RM006 in Figure 3) and several minor variant modules (not shown). The three main modules are most abundant and are different in the steps for generating two hydroxyl groups. RM004 (type 1) and RM005 (type 1a) use a dioxygenase reaction (EC 1.14.12) followed by a dehydrogenase reaction (EC 1.3.1), while RM006 (type 2) uses two steps of monooxygenase reactions (EC 1.14.13). In RM005, which often results from the preprocessing module RM003, one of the hydroxyl groups is attached together with the carboxyl group; thus, the dehydrogenase reaction is decarboxylating. Each dihydroxylation module is followed by a ring-cleavage dioxygenase reaction (EC 1.13.11) either at the meta- or ortho-position of the two hydroxyl groups.

RM008 and RM009 are the ortho- and meta-cleavage modules, respectively, for catechol and related compounds, containing the cleavage reaction and the subsequent reactions that generate TCA cycle intermediates. In the ortho-cleavage pathway, also known as beta-ketoadipate pathway,[19] catechol is converted to succinyl-CoA and acetyl-CoA, while catechol is converted to pyruvate and acetyl-CoA in the meta-cleavage pathway. It appears that the meta-cleavage is a more general processing strategy because it is a partial processing applicable to larger molecules. We defined the reaction module RM010 (Figure 3) consisting of the dihydroxylation module RM004 and a portion of the catechol meta-cleavage module RM009. A variety of chemicals including *trans*-cinnamate, *p*-cumate, ethylbenzene, styrene, and dioxin are degraded by this module resulting in 2-oxopent-4-enoate, which is further processed by RM009, and the remaining part, which may be processed differently. For example, two-ring containing biphenyl is partially processed to generate benzoate, which may further be degraded by other degradation modules.

The beta oxidation-like module RM016 mentioned above is an anaerobic version of the aromatic ring cleavage module. We have also identified an anaerobic version[20] of the preprocessing methyl oxidation module RM015 involving CoA (see Supporting Information Figure S6 for a comparison of the aerobic module RM003 and the anaerobic module RM015 for toluene degradation).

The biodegradation capacity of xenobiotic compounds, such as benzene, toluene, ethylbenzene, and xylene (BTEX compounds), is limited to specific organisms with appropriate sets of genes that are often carried by plasmids. It is therefore expected that the reaction modules identified here correspond to operon-like gene clusters in the genome. This is in fact the case. Table 5 shows the correspondence of the reaction modules (RC modules) and the KEGG modules (KO modules) for microbial biodegradation pathways, where the majority of KO modules are encoded in operon-like gene clusters (examine the Ortholog table link from each KEGG module entry, such as http://www.kegg.jp/module/M00548). For example, *Pseudoxanthomonas spadix* BD-a59[21] (KEGG organism code: psd, T01643) has preprocessing modules M00538 and M00537 for toluene and xylene, respectively, both dioxygenase and monooxygenase-catalyzed dihydroxylation modules M00551 and M00548, meta-cleave module M00569, and dioxin degradation modules M00543 and M00544, confirming its BTEX degradation potentials.

Table 5. Biodegradation Reaction Modules Corresponding to KEGG Modules

RC module	KO module	overall reaction
RM003	M00538	toluene → benzoate
	M00537	*o*-xylene → *o*-methylbenzoate
	M00419	*p*-cymene → *p*-cumate
RM004	M00547	benzoate → catechol
RM005	M00551	benzoate → catechol
RM006	M00548	benzene → catechol
RM008	M00568	catechol → 3-oxoadipate
RM009	M00569	catechol → pyruvate + acetaldehyde
RM010	M00539	*p*-cumate → 2-oxopent-4-enoate + methylpropanoate
	M00543	biphenyl → 2-oxopent-4-enoate + benzoate
RM015	M00418	toluene → benzoyl-CoA

DISCUSSION

Our analysis has shown that the metabolic network contains chemical modules of conserved reaction patterns and that these chemical reaction modules (RC modules) tend to correspond to traditional pathway modules (KO modules) and gene clusters in prokaryotic genomes. The modularity of the metabolic network has been mentioned in previous

dx.doi.org/10.1021/ci3005379 | J. Chem. Inf. Model. 2013, 53, 613−622

works,[22−25] but we believe that this is the first report on the modularity of the dual aspect of the metabolic network suggesting a coevolution of the chemical unit and the genetic unit. In contrast to the previous models of metabolic pathway evolution, such as the retrograde model,[26] the patchwork model,[27,15] and pathway duplication,[15] the chemical and genetic units are clearly defined in our approach allowing more detailed analysis. For example, as mentioned for the KO modules of M00010 and M00432 that correspond to the RC module of RM001 (Table 4), only the second and the third enzymes in *P. furiosus* are paralogs, while the reactions are similar in all the three steps. The first dissimilar enzymes may thus reflect the constraint of specific substrate recognition. The collection of KO modules and RC modules will continue to be updated in the KEGG database. Here some other reaction modules that we did not include in the Results section are briefly discussed.

We identified the reaction module RM025 for generating biogenic amines and other important metabolites from amino acids (see Supporting Information Figure S7). We already mentioned the presence and absence of a protective group in the reaction modules RM002 and RM032, respectively, where the key reaction sequence (RC00684 RC00062) is for converting a carboxyl group to an amino group following a phosphorylation step (RC00043). In contrast, RM025 contains the reverse reaction sequence (RC00062 RC00080), in which RC00080 is identical to RC00684 in the fingerprint representation, for converting an amino group to a carboxyl group following a decarboxylation step (RC00299). The overall reaction of RM025 effectively removes the main chain part of an amino acid leaving only the side chain part with a newly introduced carboxylic group. From these examples, it appears that the reaction modules contain design principles of a series of organic reactions including, for example, how to introduce a protective group, how to achieve an activated transition state, and how to increase specificity for macromolecules.

The RCLASS entries that contain the largest numbers of reactant pairs were related to phosphorylation and glycosylation. Although our approach focusing on localized structure patterns was successful to characterize possible evolution of chemical reactions with varying structural complexity, it may be necessary to consider overall structural classes, in addition to localized reaction classes, to better characterize reaction modules containing phosphorylation or glycosylation. There are cases, however, in which a special class of overall structures can be effectively extracted from the pattern of phosphorylation sequences. One such example was found in nucleotide sugar biosynthesis (map00520). Nucleotide sugars are activated forms of monosaccharides used for biosynthesis of glycans and glycosylated substances. We found two types of phosphorylation reaction modules RM022 (RC00017 RC00408 RC00002) and RM023 (RC00078 RC00002), where the first module includes an isomerization step, and the phosphate group on carbon 6 is linked to a nucleotide (see Supporting Information Figure S8).

The metabolic network may be viewed as consisting of the conserved core (primary metabolism) for maintaining life and the variable surface (secondary metabolism) for interaction with the environment. Microbial biodegradation modules (RM003 to RM016) represent convergent pathways of secondary metabolism, feeding xenobiotic compounds with varying structures to a limited number of compounds in primary metabolism. In contrast, biosynthetic pathways of secondary metabolites, especially in plants, are divergent pathways for generating a large number of structures from a limited number of compounds. We analyzed such plant biosynthetic pathways but could not identify modules as significant as the microbial degradation modules that often correspond to gene clusters in chromosomes or plasmids. Possibly, gene clusters in plants are generated from different mechanisms.[28] More importantly, the strategy to generate divergent compounds appears to be different and involves a combination of short characteristic reaction steps. The reaction module RM027 (Table 2), a hydroxylation and methylation motif, is a well-known reaction sequence involving cytochrome P450 monooxygenase. The phenylpropanoid biosynthesis pathway as represented in KEGG (map00940) is a mesh-like architecture of reactions consisting of the repeat of RM027 in the horizontal direction and the CoA involved acid−aldehyde−alcohol conversion in the vertical direction (see Supporting Information Figure S9).

The EC (Enzyme Commission) number system is a classification of enzymatic reactions, where the first three numbers in the four number system classify reactions in a hierarchical manner, namely, into class, subclass, and subsubclass. The fourth number is a sequential number given to distinguish substrate specificity. There were 4321 fully assigned EC numbers in our study, compared to 8990 reactions in KEGG. The coverage of the EC system is much more limited because the EC numbers are given only to those reactions with published records of biochemically verified enzymes. In contrast the KEGG database contains additional reactions that are supported by pathways of less characterized enzymes, in many cases the presence of genes in the genomes. We compared the EC subsubclass, which is the most detailed reaction classification in the EC system, and KEGG RCLASS (RC) that is a grouping of reactions accommodating overall structural differences of substrates. There are 249 EC subsubclasses and 2481 RC entries, and on average 13 RC entries are found for those EC subsubclasses that appear on the KEGG pathways. Obviously, the RC system presents a much finer classification of reactions than the EC system. This is shown in the list of EC subsubclasses that contain the largest numbers of RC entries (Table 6), where the most notable example is monooxygenases. In the EC system, monooxygenases (1.14) are further classified based on the cofactors (1.14.13, 1.14.14, 1.14.15, etc.), but the distinction is not made

Table 6. List of EC Numbers Containing the Largest Numbers of RC Entries

EC subsubclass	no. of RC entries	enzymes involved
1.14.13	150	monooxygenases
4.2.1	129	hydratases/dehydratases, terpene cyclases (hydrating)
1.1.1	117	alcohol:NAD(P)+ dehydrogenases
4.2.3	86	phospho-lyases, terpene cyclases (diphosphate-eliminating)
1.14.14	83	monooxygenases
2.1.1	79	methyltransferases
1.13.11	72	dioxygenases
1.3.1	68	saturases/desaturases
4.1.1	65	carboxylases/decarboxylases
2.5.1	60	prenyltransferases, 1-carboxyvinyltransferases, aminocarboxyethyltransferases, aminocarboxypropyltransferases, adenosyltransferases

dx.doi.org/10.1021/ci3005379 | J. Chem. Inf. Model. 2013, 53, 613−622

for reaction mechanisms. For example, CYP monooxygenases may further be classified into hydroxylases, epoxidases (epoxygenases), Baeyer−Villiger monooxygenases, and so on (see Supporting Information Figure S10). It is interesting to note that the full EC number 1.14.14.1 (single reaction with the same substrate specificity in the EC system) corresponds to 75 RC entries (different reactions in the RC system).

There is an important practical value of the dual aspect of metabolism. Knowledge of genes and proteins can be used to make predictions about chemical compounds and reactions and, conversely, knowledge of chemical compounds and reactions can be used to make predictions about genes and proteins. The present analysis has already resulted in the improvement of the KEGG pathway maps for chlorocyclohexane and chlorobenzene degradation (map00361), benzoate degradation (map00362), toluene degradation (map00623), and xylene degradation (map00622) in terms of both identifying additional reactions and improving genome annotations. In the KEGG database, the repertoire of genes in all available complete genomes are categorized by the KEGG Orthology (KO) system, an attempt to understand the genomic space of life. Here the RC system has been introduced to categorize all known reactions, an attempt to understand the chemical space of life and the surrounding environment. We are trying to establish empirical relationships between the KO system and the RC system, especially in terms of the modular structures of the metabolic network represented by the KO modules and the RC modules. The logic of chemical reactions is so straightforward, compared to, for example, protein−protein interactions or protein−DNA interactions, that these empirical relationships can be safely extended to make predictions about metabolism resulting either from new genes or from new reactions. This type of predictive annotation is already implemented in the KEGG database toward better understanding of metabolic features encoded in the genome.

■ ASSOCIATED CONTENT

Supporting Information

KEGG atom types, RDM patterns, fingerprint representation, and similarity scoring of RC entries (Methods); tricarboxylic acid pathway (Figure S1); glucosinolate biosynthesis pathway (Figure S2); carboxyl to amino conversion (Figure S3); fatty acid biosynthesis (Figure S4); beta oxidation (Figure S5); methyl to carboxyl conversion on aromatic ring (Figure S6); amino to carboxyl conversion (Figure S7); nucleotide sugar biosynthesis (Figure S8); phenylpropanoid biosynthesis (Figure S9); and classification of monooxygenases (Figure S10). This information is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +81-774-38-4521. Fax: +81-774-38-3269. E-mail: kanehisa@kuicr.kyoto-u.ac.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Nicolas Joannin for critical reading of the manuscript. This work was supported by the Japan Science and Technology Agency. Computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

■ REFERENCES

(1) Bono, H.; Ogata, H.; Goto, S.; Kanehisa, M. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Res.* **1998**, *8*, 203−210.

(2) Galperin, M. Y.; Koonin, E. V. Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. *Genetica* **1999**, *106*, 159−170.

(3) Dandekar, T.; Schuster, S.; Snel, B.; Huynen, M.; Bork, P. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.* **1999**, *343*, 115−124.

(4) Forst, C. V.; Schulten, K. Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. *J. Comput. Biol.* **1999**, *6*, 343−360.

(5) Ogata, H.; Fujibuchi, W.; Goto, S.; Kanehisa, M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.* **2000**, *28*, 4021−4028.

(6) Tohsato, Y.; Matsuda, H.; Hashimoto, A. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2000**, *8*, 376−383.

(7) Pinter, R. Y.; Rokhlenko, O.; Yeger-Lotem, E.; Ziv-Ukelson, M. Alignment of metabolic pathways. *Bioinformatics* **2005**, *21*, 3401−3408.

(8) Wernicke, S.; Rasche, F. Simple and fast alignment of metabolic pathways by exploiting local diversity. *Bioinformatics* **2007**, *23*, 1978−1985.

(9) Tohsato, Y.; Nishimura, Y. Reaction similarities focusing substructure changes of chemical compounds and metabolic pathway alignments. *Inf. Media Technol.* **2009**, *4*, 390−399.

(10) Ay, Y.; Kellis, M.; Kahveci, T. SubMAP: aligning metabolic pathways with subnetwork mappings. *J. Comput. Biol.* **2011**, *18*, 219−235.

(11) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2012**, *40*, D109−D114.

(12) 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.

(13) McDonald, A. G.; Boyce, S.; Tipton, K. F. ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.* **2009**, *37*, D593−D597.

(14) 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.

(15) Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **1976**, *30*, 409−425.

(16) Horie, A.; Tomita, T.; Saiki, A.; Kono, H.; Taka, H.; Mineki, R.; Fujimura, T.; Nishiyama, C.; Kuzuyama, T.; Nishiyama, M. Discovery of proteinaceous N-modification in lysine biosynthesis of Thermus thermophilus. *Nat. Chem. Biol.* **2009**, *5*, 673−679.

(17) Maeder, D. L.; Weiss, R. B.; Dunn, D. M.; Cherry, J. L.; González, J. M.; DiRuggiero, J.; Robb, F. T. Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences. *Genetics* **1999**, *152*, 1299−1305.

(18) Pelletier, D. A.; Harwood, C. S. 2-Hydroxycyclohexanecarboxyl coenzyme A dehydrogenase, an enzyme characteristic of the anaerobic benzoate degradation pathway used by *Rhodopseudomonas palustris*. *J. Bacteriol.* **2000**, *182*, 2753−2760.

(19) Harwood, C. S.; Parales, R. E. The beta-ketoadipate pathway and the biology of self-identity. *Annu. Rev. Microbiol.* **1996**, *50*, 553−590.

(20) Rabus, R.; Kube, M.; Heider, J.; Beck, A.; Heitmann, K.; Widdel, F.; Reinhardt, R. The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1. *Arch. Microbiol.* **2005**, *183*, 27−36.

(21) Lee, S. H.; Jin, H. M.; Lee, H. J.; Kim, J. M.; Jeon, C. O. Complete genome sequence of the BTEX-degrading bacterium *Pseudoxanthomonas spadix* BD-a59. *J. Bacteriol.* **2012**, *194*, 544.
(22) Papin, J. A.; Reed, J. L.; Palsson, B. O. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. *Trends Biochem. Sci.* **2004**, *29*, 641−647.
(23) Ravasz, E.; Somera, A. L.; Mongru, D. A.; Oltvai, Z. N.; Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science* **2002**, *297*, 1551−1555.
(24) Schuster, S.; Pfeiffer, T.; Moldenhauer, F.; Koch, I.; Dandekar, T. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics* **2002**, *18*, 351−361.
(25) Yamada, T.; Kanehisa, M.; Goto, S. Extraction of phylogenetic network modules from the metabolic network. *BMC Bioinf.* **2006**, *7*, 130.
(26) Horowitz, N. H. On the evolution of biochemical synthesis. *Proc. Natl. Acad. Sci USA* **1945**, *31*, 153−157.
(27) Ycas, M. On earlier states of the biochemical system. *J. Theor. Biol.* **1974**, *44*, 145−160.
(28) Chu, H. Y.; Wegel, E.; Osbourn, A. From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *Plant J.* **2011**, *66*, 66−79.