代謝経路からの酵素反応連続パターンの検出

Extraction of tandem enzymatic reaction patterns from metabolic pathways

京都大学化学研究所バイオインフォマティクスセンター
化学生命科学　　武藤　愛

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 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. 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 measure for systematic analysis, for example, comparison of genomic diversity of enzyme genes and chemical diversity of enzyme-catalyzed reactions.

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 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.

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.

Reactant pairs. The KEGG REACTION database contains all known enzymatic reactions taken from the Enzyme Nomenclature and also from the metabolic pathway section of the KEGG PATHWAY database. The KEGG release that we used contains 8,990 reactions including 4,321 Enzyme Nomenclature reactions. Among them 6,238 reactions including 2,595 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. 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 (described below) changes at the reaction center (R), the difference substructure (D), and the matching substructure (M) atoms. The RDM notation for the pair of reactants A and B is as follows:

$$\mathrm{RDM}(A, B) = \mathrm{R}A\text{-}\mathrm{R}B : \mathrm{D}A\text{-}\mathrm{D}B : \mathrm{M}A\text{-}\mathrm{M}B$$

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

RDM = N1a-N1b:*-C5a:C1b-C1b

The KEGG atom type generally consists of three characters. The first (upper case letter) indicates the atomic species, the second (numeral) represents the pre-defined class of atomic bonding for each atomic species, and the third (lower case letter) represents the pre-defined 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.

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. 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 2,481 RCLASS entries in this study.

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 pi bond, a carbon atom in 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 Supplementary material 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.

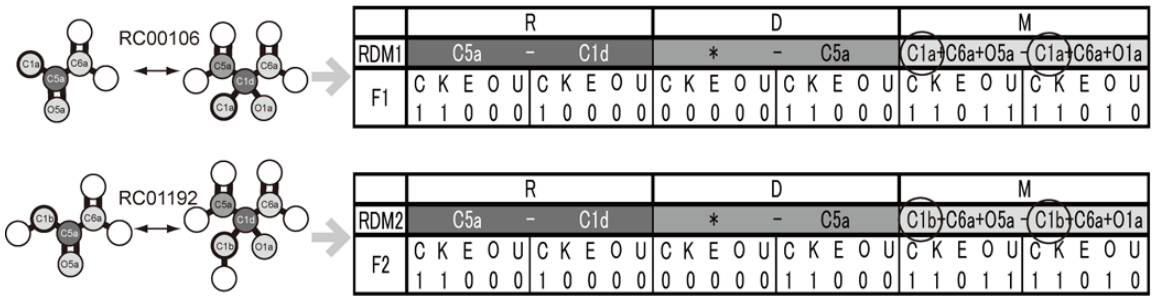

Figure 1. Fingerprint representation of the RDM pattern.

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

$$S(RC_1, RC_2) = w_R J_a(v_{R1}, v_{R2}) + w_D J_a(v_{D1}, v_{D2}) + w_M J_a(v_{M1}, v_{M2})$$

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 Supplementary material 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, 2,481 RCLASS entries were grouped into 376 similarity groups with more than one member and the remaining 1,190 singletons.

RESULTS and DISCUSSION

Extraction of conserved RCLASS sequence patterns. Based on 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 Supplementary material for more details). Known metabolic pathways in the KEGG PATHWAY database are split into all possible subsequences of 2 to 8 consecutive reactions. The pathways involving branches are split into all combinations of linear reaction sequences.

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.

General characteristics of reaction modules. The list of manually refined reaction modules is 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 of 21 RC modules, and all the KO modules except one contained operon-like gene clusters in some genomes. Here we report detailed analysis of the reaction modules for 2-oxocarboxylic acid chain extension and modification.

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 tricarboxylic acids in the TCA cycle (map00020 in KEGG), where 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. Furthermore, in the biosynthesis pathway of glucosinolates (map00966), which are plant

Table 1. The number of conserved RCLASS sequence patterns found in the KEGG metabolic pathways.

Length	#of conserved patterns	# of reactions included	Coverage*
2	928	3,479	0.599
3	770	2,503	0.431
4	534	1,662	0.286
5	338	1,074	0.185
6	218	765	0.132
7	140	527	0.091
8	88	399	0.069
Total	3,016		

* The ratio to 5,805 reactions, the total number of reactions with RC assignment in the KEGG pathways.

secondary metabolites, a six tandem repeat of RM001 is found from 2-oxo-4-methylthiobutanoate to 2-oxo-10-methylthiodecanoate.

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).

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).

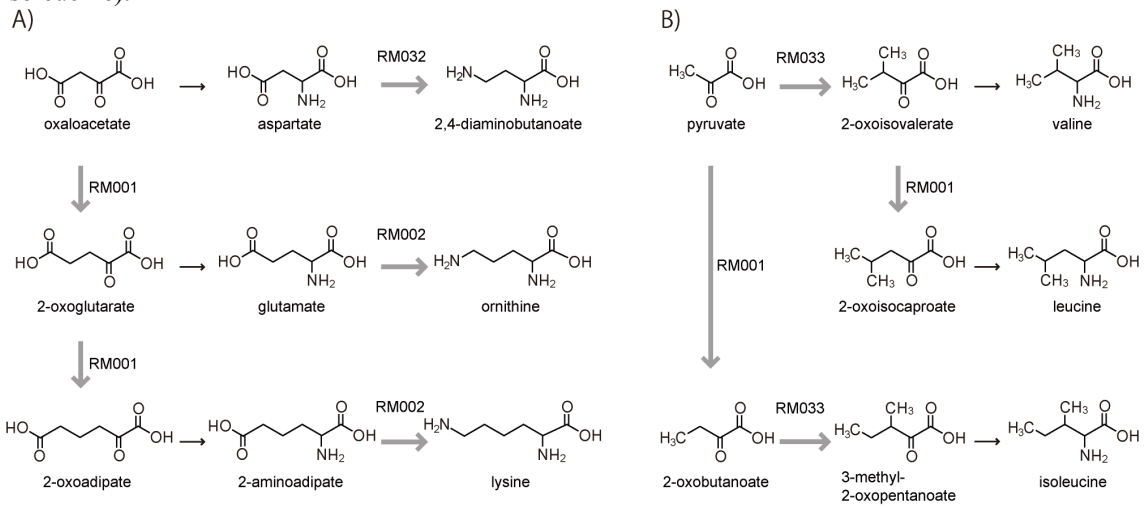

Figure 2. The architecture of reaction modules consisting of 2-oxocarboxylic acidchain extension and modification

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.We found, for example, that the RC module RM001 coincided well with the KO modules M00010, M00432 and M00535. As shown in Table 3, 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. 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 3. Reaction modules corresponding to enzyme gene clusters.

RC	Overall reaction	KO	Gene cluster example
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
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