GENERALIZED REACTION PATTERNS FOR PREDICTION OF UNKNOWN ENZYMATIC REACTIONS

YUGO SHIMIZU
shimizu@kuicr.kyoto-u.ac.jp

MASAHIRO HATTORI
hattori@kuicr.kyoto-u.ac.jp

SUSUMU GOTO
goto@kuicr.kyoto-u.ac.jp

MINORU KANEHISA
kanehisa@kuicr.kyoto-u.ac.jp

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

Prediction of unknown enzymatic reactions is useful for understanding biological processes such as reactions to external substances like endocrine disrupters. To create an accurate prediction, we need to define a similarity measure in the reaction. We have developed the KEGG RPAIR database which is a collection of chemical structure transformation patterns, called RDM patterns, for substrate-product pairs of enzymatic reactions. In this study, we compared RDM patterns with EC numbers which are the well-known hierarchical classification scheme for enzymes. Additionally, we performed hierarchical clustering of RDM patterns using the information stating whether each sub-subclass of EC has a particular RDM pattern or not. To represent the variation of RDM patterns in a cluster, we generalized RDM patterns in the same cluster using the hierarchy of KEGG Atomtypes, which are the components of RDM patterns. Using this generalized pattern, we can predict which cluster includes a given RDM pattern even if the reaction of the pattern has not been assigned any EC numbers. Thus we will be able to define the similarity between enzymatic reactions by using this cluster information.

Keywords: EC number; KEGG RPAIR; classification of enzymes; enzymatic reaction

1. Introduction

Recently, a large amount of biochemical information as well as genomic information and chemical information has become available [5, 6]. For example, in the KEGG LIGAND database, much information about biochemical small molecules, biochemical reactions, enzymes, glycans, and drugs are available [1, 12]. Here enzymes are proteins that catalyze the biochemical reactions; however there are lots of enzymes whose function have yet to be unveiled. This causes missing enzymes in metabolic pathways and many unknown reactions should be characterized. Thus, the computational prediction of unknown enzymatic reactions may be useful for understanding the biological processes such as xenobiotics biodegradation: reactions to external substances like endocrine disrupter [2, 7, 8]. To improve the accuracy of prediction, we need to better systematize the reaction mechanisms of known enzymatic activities and to define an appropriate measure of similarity among the enzymatic reactions for further analysis. To achieve these objectives, we performed comprehensive analyses using the EC classification and KEGG RPAIR database.

The EC (Enzyme Commission) number is a well-known classification scheme for enzymes [9, 11]. In EC classification, enzymes are hierarchically classified by types of catalyzed reactions and their substrates and products. Each EC number consists of the letters "EC" followed by four numbers separated by periods (e.g. EC 1.1.1.1). The first, second, and third numbers are called class, subclass, and sub-subclass respectively. The fourth number represents the substrate specificity. The EC numbers have been utilized for many computational applications such as classification or prediction of enzymatic reactions. However, there are also some problems in EC classification. The EC numbers are classified manually, based on published experimental data, by the IUPAC-IUBMB Joint Commission on Biochemical Nomenclature. This requirement of published articles leaves many reactions unclassified. Additionally, the structural transformation between single compounds pair is unclear since EC represents the relationships between multiple substrates and multiple products. In order to avoid these problems, we have developed the KEGG RPAIR database that is a collection of chemical structure transformation patterns, called RDM patterns, for every substrate-product pairs of enzymatic reactions [4]. In this study, we compared the RDM patterns with EC numbers and performed hierarchical clustering of the RDM patterns using the information whether each sub-subclass of EC has the RDM pattern or not. To represent the variation of the RDM patterns in a cluster, we introduced the generalized RDM patterns in the same cluster using the hierarchy of KEGG Atomtypes, which are the components of the RDM patterns.

2. Materials and Methods

2.1. *KEGG LIGAND database*

KEGG LIGAND is a composite database which contains various databases about biochemical compounds. In this study we have used ENZYME, REACTION, and RPAIR from the KEGG LIGAND database (as of 2008/05/13). ENZYME (4976 entries) is a database of EC numbers and contains names of enzymes, catalyzed reactions, genes, as well as other types of information. REACTION (7567 entries) is a database of all biochemical reactions that are included in ENZYME or appear on KEGG metabolic pathways. RPAIR (8706 entries) is a database of chemical structure transformation patterns, called RDM patterns, for every substrate-product pair (reactant pair) in REACTION.

2.2. *RDM pattern*

2.2.1. *KEGG RPAIR database*

Each entry in RPAIR contains the alignment of atoms between the substrate-product pairs and the structural transformation pattern called RDM pattern. In general, one enzymatic reaction contains multiple substrates and multiple products, which result in

multiple pairs. Here each pair of chemical compounds should be distinguished by its biochemical role under the reaction, and in the RPAIR database five types of such roles have been available with the annotated labels, “main”, “cofac”, “leave”, “ligase” and “trans”, which are exemplified in Figure 1. In this study, to reduce the noise of poorly characterized pairs we used only the main type which corresponds to a major component of pairs in each reaction.

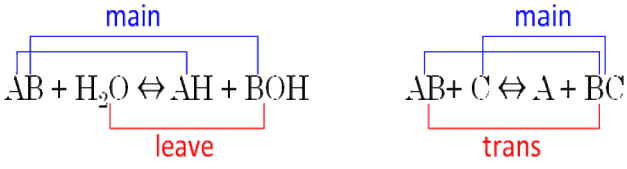

Fig. 1. Examples of substrate-product pairs and their assigned types. In the left example, both the pair (AB, AH) and pair (AB, BOH) are classified as the main type and the other pair (H_2O, BOH) is defined as the leave type. In the right example, there are also two main types and the trans type is assigned the last one. In any cases, the hydrogen atoms are not considered.

Table 1. Definition of KEGG Atomtypes.

(extracted from http://www.genome.jp/kegg/reaction/KCF.html)

Atom	Atom class	Description	Atomtype	Description
C	C1	alkane	C1a	R-CH_3
			C1b	R-CH_2-R
			C1c	R-CH(-R)-R
			C1d	R-C(-R)$_2$-R
			C1x	ring-CH_2-ring
			C1y	ring-CH(-R)-ring
			C1z	ring-C(-R)$_2$-ring
	C2	alkene	C2a	R=CH_2
			C2b	R=CH-R
			C2c	R=C(-R)$_2$
			C2x	ring-CH=ring
			C2y	ring-C(-R)=ring or ring-C(=R)-ring
O	O1	single bond	O1a	R-OH
			O1b	N-OH
			O1c	P-OH
			O1d	S-OH
			O2a	R-O-R
			O2b	P-O-R
			O2c	P-O-P
			O2x	ring-O-ring

2.2.2. *KEGG Atomtype*

In KEGG RPAIR, all atoms are represented by KEGG Atomtypes, which have been hierarchically defined by the physicochemical environment of atoms. Mostly, atomtypes are represented as three letter codes as shown in Table 1. The first letter indicates the atomic species, the second indicates information about the atomic bonds, and the third

indicates the information of the substituted groups. In particular, the second level of hierarchy in KEGG Atomtypes is called the atom class. For example, "C" is the carbon atom itself, the atom class "C1" represents the carbon atom observed in alkanes and the atomtype "C1a" represents the carbon atom which connects to another carbon atom and three hydrogen atoms. There are 68 atomtypes in RPAIR database and a portion of them is shown in Table 1.

2.2.3. *RDM pattern*

An RDM pattern is defined as a set of KEGG Atomtype changes at the reaction center (R), the difference region (D), and the matched region (M) for each reactant pair (Fig. 2). R atoms are boundary atoms between the matched regions and the unmatched regions. D atoms are next to the reaction center (R atoms) in the unmatched regions. M atoms are adjacent to the R atoms in the matched regions. In most cases R, D, and M atoms are all single pairs and the RDM pattern is represented as "R_1-R_2:D_1-D_2:M_1-M_2" (Fig. 2). Multiple pairs in D or M atoms can be considered and are represented by concatenating all atomtypes using "+", and multiple pairs in R are represented by multiple RDM patterns in which R atoms are a single pair. The asterisk "*" in the RDM patterns indicates that there is no atom or it is only a hydrogen atom. The structural transformation between single compounds pair is now clear since each entry of the RPAIR database is a binary pair. Also the RDM pattern represents the transformational pattern around the reaction center. Hence it can be assumed that the RDM patterns may basically reflect the reaction mechanism at the site where each enzyme catalyzes. RDM patterns are generated first computationally by the chemical structure comparison program SIMCOMP, followed by manual curation [3]. There were 2401 kinds of the RDM patterns in RPAIR.

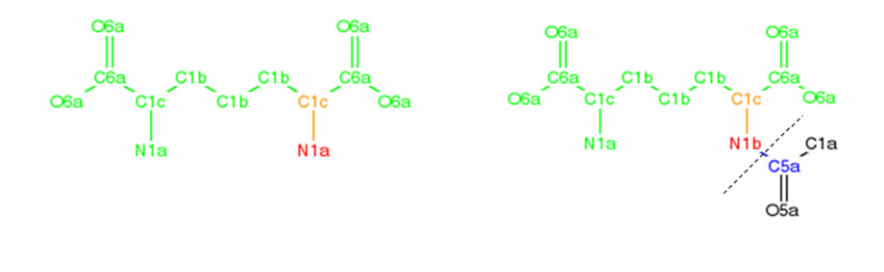

Fig. 2. Examples of a substrate-product pair and its RDM pattern. The red colored atoms (N1a and N1b in the boundary of the dashed line) are R, the blue and the yellow atoms are D (C5a connected to N1b) and M (C1b connected to R atoms), respectively. The rest of the matched region is depicted by green color.

2.3. *EC-RDM dot matrix*

All EC numbers and corresponding RDM patterns were extracted from the databases. Then, the EC-RDM dot matrix was created to overview the relationship between the EC classification and the RDM patterns. The row of the matrix corresponds to sub-subclasses of EC numbers, and the column of the matrix corresponds to the RDM patterns. The characteristic relationship between EC sub-subclasses and RDM patterns in the matrix is shown in the Result section.

2.4. *Hierarchical clustering*

After obtaining the EC-RDM dot matrix, we performed a hierarchical clustering of the RDM patterns, using the information whether each EC sub-subclass has a particular RDM pattern or not. The distance (D) between two RDM patterns (RDM_1 and RDM_2) can be formulated as follows:

$$D(RDM_1, RDM_2) = 1 - Tc(V(RDM_1), V(RDM_2)) \quad (1)$$

where $V(RDM_1)$ and $V(RDM_2)$ are respective bit vectors of the RDM patterns RDM1 and RDM2, and each element of a vector corresponds to the existence (1) or nonexistence (0) of each sub-subclass of EC. Tc indicates the Tanimoto coefficient which is defined as follows:

$$Tc(x,y) = \frac{\text{The number of bits where } x_i = 1 \text{ and } y_i = 1}{\text{The number of bits where } x_i = 1 \text{ or } y_i = 1} \quad (2)$$

where $\{x_i\}$ and $\{y_i\}$ are bit vectors [10]. We used the average linkage method for the hierarchical clustering.

2.5. *Generalization of RDM patterns*

Using the cluster information obtained in the above section, we constructed the generalized patterns of the RDM patterns to represent the variation of the RDM patterns in each cluster. We implemented an algorithm that compares character strings of the RDM patterns in the same cluster to generate their generalized pattern. In this generalization process, the hierarchy of KEGG Atomtypes (atom species, atom class, and atomtype) is used. The detailed procedure of generalizing two RDM patterns, RDM_1 and RDM_2, is described as follows:

Step 1: All possible representations of RDM_1 are generated and stored in $\{RDM_1\}$.
Step 2: The following procedures (2-2) are performed for each RDM_{1i} of the set $\{RDM_1\}$.

Step 2-2: RDM_{1i} is separated into R_{1i}, D_{1i}, and M_{1i}. RDM_2 is also separated into R_2, D_2, and M_2. Then, R_{1i} and R_2, D_{1i} and D_2, and M_{1i} and M_2 are compared respectively. When multiple atoms are incorporated into each D or M representation, they are compared at the corresponding position of atoms. That is, when comparing D_{1i} (= D^1_{1i}+D^2_{1i}) with D_2 (= D^1_2+D^2_2), the comparison is done between D^1_{1i} and D^1_2 and between D^2_{1i} and D^2_2.

Step 3: The most matched case is selected and the generalized pattern is generated.

The priority of the matching the atom representations when comparing KEGG Atomtypes in Step 2-2 is shown in Table 2. Generalized patterns are made via following conditions. The example of generalization is also shown in Table 2 and Fig. 3.

i) The parts which have complete match in Step 2-2 are output directly.
ii) The parts which have match at the atom class level or atom species level in Step 2-2 are substituted by the atom class or atom species respectively.
iii) The parts which have no match in Step 2-2 are substituted by both components separated by comma and in parentheses.

Table 2. Definition of the priority in the atomtype comparison and examples of generalization between atomtypes.

Priority	Description	Example of generalization	
		Original atomtypes	Generalized pattern
1	Complete match	P1b and P1b	P1b
2	Matching at the atom class level	O2c and O2b	O1
3	Matching at the atom species level	O1c and O3b	O
4	No match in comparison	P1b and C1b	(P1b,C1b)

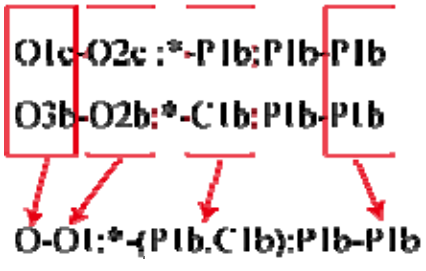

Fig. 3. An example of generalization.

3. Results

3.1. *Relationship between EC sub-subclasses and RDM patterns*

There were 3116 EC numbers (195 EC sub-subclasses) which correspond to at least one RDM pattern (1571 main types). Fig. 4 shows the EC-RDM dot matrix.

Some RDM patterns correspond to many sub-subclasses of EC numbers. For examples, the RDM pattern "O1c-O2c:*-P1b:P1b-P1b" in the box A in Fig. 4 corresponds to 25 sub-subclasses of EC numbers and "S1a-S2a:*-C5a:C1b-C1b" in the box B corresponds to 14 sub-subclasses of EC numbers. These patterns are found in reactions such as the hydrolysis of ATP and the formation of a thioester bond respectively. These reactions are most significant and can be observed extensively in biochemical reactions since they are frequently used as the energy source of other reactions.

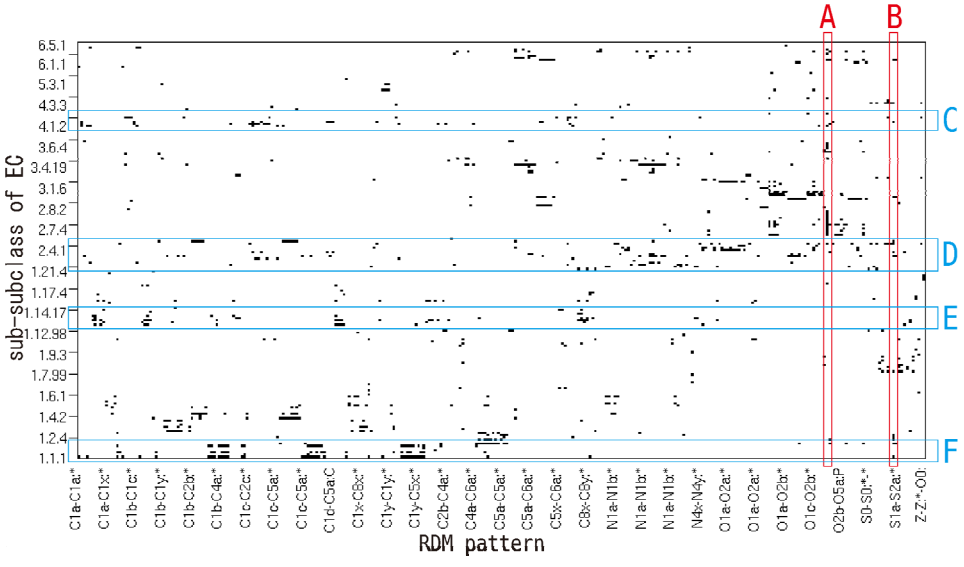

Fig. 4. The EC-RDM dot matrix. The RDM patterns corresponding to at least 2 sub-subclasses of EC are shown because of simplicity.

Some EC sub-subclasses found in the boxes C, D, E and F correspond to many RDM patterns. For example, EC 1.1.1 (box F), EC 4.2.1 (box C) and EC 2.5.1 (box D) correspond to 98, 83 and 72 RDM patterns respectively. The number of EC numbers within a certain EC sub-subclass is different depending on the sub-subclass. Therefore we compared the number of the RDM patterns within each sub-subclass of EC with that of enzymes included in each sub-subclass of EC (Fig. 5). As seen in the Fig. 5, the

variation of the RDM patterns in sub-subclass of EC almost depends on the variation of the 4th number of EC except EC 2.7.11 and EC 3.6.3 which have only one RDM pattern.

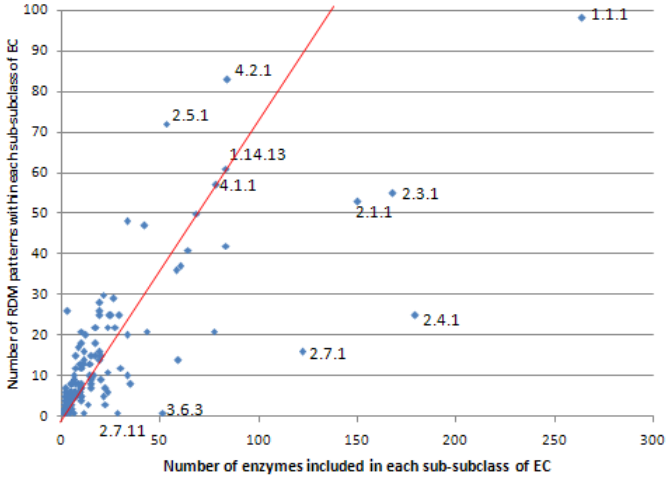

Fig. 5. Relationship between the numbers of enzymes included in each sub-subclass of EC and the number of the RDM patterns within each sub-subclass of EC.

3.2. *Hierarchical clustering and generalization of RDM patterns*

We performed the hierarchical clustering of the RDM patterns by using the information of existence or nonexistence of sub-subclasses of EC numbers. A part of the resulting cluster is shown in Fig. 6. It is obvious that the RDM patterns belonging to the same cluster consist of similar character strings and that the diversity of KEGG Atomtypes in the RDM patterns are considerably low. On the other hand, KEGG Atomtypes involved with other clusters are much different from each other, and the difference of such atomtype representations becomes larger and larger at the higher-level of the hierarchy in the clustering tree.

Fig. 6. A part of the clustering tree of the RDM pattern. Only RDM patterns corresponding to at least 2 sub-subclasses of EC are shown because of simplicity. The RDM patterns are shown at the leaves of the clustering tree with their corresponding EC sub-subclasses. The full image of the resulting tree is available at the following URL. http://web.kuicr.kyoto-u.ac.jp/supp/shimizu/ibsb2008/

Next, we performed the generalization of the RDM patterns. The generalization starts at the lowest level of the cluster tree, that is the leaf of the tree, then grows up to higher level of hierarchy, and end at the highest level of cluster, that is the root of the tree. The generalization process is exemplified in Fig. 7. In particular case of this figure, the generalized pattern of RDM at the highest level is O1-O2a:*-C1:C-C, which can contain all of the RDM patterns within the whole clusters. We applied this generalization to all clusters. The generalized pattern of RDM which correspond to at least 2 sub-subclasses of EC had comparatively simple forms, however that of RDM which correspond to only 1 sub-subclass of EC tended to have somewhat complicated forms (e.g. (N,C,O1b,)-(N,C,O7a):*-(C,O):(N1,C,O6)) even if the distance between clusters was equal to 0. This is because the clusters tend to have many RDM patterns and the diversity within a cluster itself becomes larger.

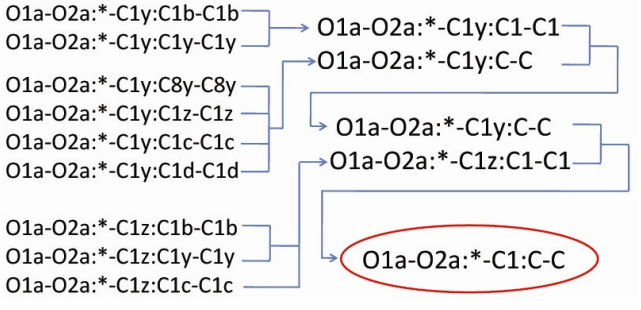

Fig. 7. An example of the result of the RDM generalization.
The clustering tree of the RDM pattern is the same as that of Figure 6.

4. Discussion

In this study, we have systematized the reaction mechanisms based on the EC classification by hierarchically clustering the RDM patterns. We could successfully represent the variation of reactions by using the generalized pattern of RDM. For example, a generalized pattern "N1b-N1a:C2-*:C1b-C1b" can represent five possible patterns, since the atom class "C2" contains the following five atomtypes: C2a, C2b, C2c, C2x, and C2y. Because a generalized pattern is generated by the patterns in the same cluster in which combinations of corresponding EC numbers are similar, variations in generalized patterns indicate the possible reaction patterns in some EC numbers. Using this generalization we will be able to calculate which cluster includes a given RDM pattern even if the relevant reaction has never been assigned to any EC numbers. Then we will be able to define the similarity measure between known enzymatic reactions (which are found in the database) and unknown enzymatic reactions and consequently we may improve the accuracy of the prediction of unknown enzymatic reactions. For example, we have developed the e-zyme system, which can automatically assign the EC number to a given compound pair by using the RDM patterns. Incorporating the

generalization and the similarities of the RDM patterns in the EC assignment process of the e-zyme, we will be able to improve its accuracy rate.

Acknowledgments

We would like to thank J.B. Brown for the proofreading of our manuscript. This work was supported in part by a grant-in-aid for scientific research on the priority area "Comprehensive Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan, 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.

References

[1] Goto, S., Okuno, Y., Hattori, M., Nishioka, T. and Kanehisa, M., LIGAND: database of chemical compounds and reactions in biological pathways, *Nucleic Acids Res.*, 30(1): 402-404, 2002

[2] Hendry L.B., Roach L.W. and Mahesh V.B., Multidimensional screening and design of pharmaceuticals by using endocrine pharmacophores, *Steroids*, 64(9): 570-575, 1999.

[3] Hattori, M., Okuno, Y., Goto, S., and 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.*, 125(39):11853-11865, 2003.

[4] Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M., Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions, *J. Am. Chem. Soc.*, 126(50):16487-16498, 2004.

[5] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M., From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res.,* 34:D354-357, 2006.

[6] Kanehisa, M. and Bork, P. Bioinformatics in the post-sequence era, *Nature Genetics*, 33:305-310, 2003.

[7] Oh, M., Yamada, T., Hattori, M., Goto, S., and Kanehisa, M., Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways, *J. Chem. Inf. Model.*, 47(4):1702-1712, 2007.

[8] Phillips K.P. and Foster W.G., Key Developments in Endocrine Disrupter Research and Human Health, *J. Toxicol. Environ. Health B Crit. Rev.*, 11(3-4):322-344, 2008.

[9] Webb, E.C. and International Union of Biochemistry and Molecular Biology. Nomenclature Committee, *Enzyme Nomenclature*, Academic Press, San Diego, California, 1992.

[10] Willett P., Barnard J. M., and Downs G. M. Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci*., 38: 983-996, 1998.

[11] http://www.chem.qmul.ac.uk/iubmb/enzyme/

[12] http://www.genome.jp/kegg/ligand.html