JACS
ARTICLES
Published on Web 11/26/2004

Computational Assignment of the EC Numbers for Genomic-Scale Analysis of Enzymatic Reactions

Masaaki Kotera, Yasushi Okuno,[†] Masahiro Hattori, Susumu Goto, and Minoru Kanehisa*

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

Received June 8, 2004; E-mail: kanehisa@kuicr.kyoto-u.ac.jp

Abstract: The EC (Enzyme Commission) numbers represent a hierarchical classification of enzymatic reactions, but they are also commonly utilized as identifiers of enzymes or enzyme genes in the analysis of complete genomes. This duality of the EC numbers makes it possible to link the genomic repertoire of enzyme genes to the chemical repertoire of metabolic pathways, the process called metabolic reconstruction. Unfortunately, there are numerous reactions known to be present in various pathways, but they will never get EC numbers because the EC number assignment requires published articles on full characterization of enzymes. Here we report a computerized method to automatically assign the EC numbers up to the sub-subclasses, i.e., without the fourth serial number for substrate specificity, given pairs of substrates and products. The method is based on a new classification scheme of enzymatic reactions, named the RC (reaction classification) number. Each reaction in the current dataset of the EC numbers is first decomposed into reactant pairs. Each pair is then structurally aligned to identify the reaction center, the matched region, and the difference region. The RC number represents the conversion patterns of atom types in these three regions. We examined the correspondence between computationally assigned RC numbers and manually assigned EC numbers by the jackknife cross-validation test and found that the EC sub-subclasses could be assigned with the accuracy of about 90%. Furthermore, we examined the correlation with genomic information as represented by the KEGG ortholog clusters (OC) and confirmed that the RC numbers are correlated not only with elementary reaction mechanisms but also with protein families.

Introduction

A major challenge in the postgenomic era is to elucidate cellular functions as behaviors of complex interaction systems. The advancements of high-throughput experimental technologies and computational methods increasingly allow us to accumulate and analyze large-scale data in the genome, in the transcriptome, and in the proteome toward understanding such systems.[1] Here we focus our attention on the metabolome, which in our definition is a multiple system consisting of genes, enzymes, chemical compounds, and biological macromolecules including glycans, lipids, and nonribosomal peptides. The metabolic reconstruction or the mapping of enzyme genes in the genome to the primary pathways of intermediary metabolism is relatively straightforward now with the help of the pathway databases such as KEGG.[2,3] However, it is often the case that a large fraction of putative enzyme genes remains to be characterized both as proteins with specific catalytic functions and as pathways of secondary metabolism. When the intermediary metabolism is viewed as the core of the metabolome, the secondary metabolism forms the shell that is in contact with the environments. The metabolome is a dynamic interaction system where varying chemical environments continuously affect genomic contents and vice versa. We are thus developing computational methods to uncover empirical relations between genomic and chemical information in the metabolome.

The basis of linking genomics and chemistry is the EC (Enzyme Commission) numbers.[4] The EC numbers represent enzymatic reactions (chemical information), but they are also utilized as identifiers of enzymes and enzyme genes (genomic information). The assignment of the EC numbers is performed manually, based on published experimental data on individual enzymes, by the Joint Commission on Biochemical Nomenclature (JCBN) of the International Union of Biochemistry and Molecular Biology (IUBMB) and the International Union of Pure and Applied Chemistry (IUPAC). Unfortunately, however, a requirement of published articles on individual enzymes leaves many reactions unassigned, such as reactions known to be present in pathways and reactions inferred from chemical compounds. In fact, the majority of the enzymes in the secondary metabolism are unlikely to receive EC numbers

[†] Present address: Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan.

[1] Kanehisa, M.; Bork, P. Bioinformatics in the post-sequence era. *Nat. Genet.* **2003**, *33*, 305−310.

[2] Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **1997**, *13*, 375−376.

[3] Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **2004**, *32*, D277−D280.

[4] 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, CA, 1992.

10.1021/ja0466457 CCC: $27.50 © 2004 American Chemical Society

because each reaction step will never be fully characterized in a traditional way.

To supplement the current EC numbers, which must be viewed as a curated set of well-characterized enzymatic reactions, we have developed an automatic EC number assignment system, which can be applied to any reaction fully or partially characterized. Our assignment system is based purely on chemical knowledge, without any use of protein sequence or other information on enzymes. Given pairs of substrates and products, the system assigns an RC (reaction classification) number to each pair using the chemical structure alignment method developed before,[5] and the best matching EC sub-subclass is determined based on the current set of RC−EC correspondences. Naturally, our method uncovers irregularities in the current EC numbers. There are cases where enzymes in the same EC sub-subclass consist of multiple types of reactions and where the same reaction is given multiple EC sub-subclasses. The EC numbering system is an accumulation of human knowledge, sometimes reflecting opinions and compromises. In contrast, the RC numbering system is computationally derived based on chemical structure comparisons, in a similar spirit with the protein classification that is performed computationally based on sequence or 3D structure comparisons. This way, we hope to extract common characteristics in different reactions that have not been apparent before and also to better understand relations between reaction mechanisms end enzyme families.

Materials and Methods

Data Sets. The REACTION section of the KEGG LIGAND database[3,6] contains individual reactions of all the EC numbers, where a single EC number may correspond to multiple reactions. We used 5227 reactions in 3254 EC numbers, from which we extracted 8605 reactant pairs. As described below in more detail, reactant pairs are manually selected substrate−product pairs from the reaction formulas, and each reactant pair is subjected to a chemical structure comparison in order to assign an RC (reaction classification) number.

Reactant Pairs. Reactant pairs are defined as pairs of compounds that have atoms or atom groups in common on two sides of a reaction. For example, alcohol dehydrogenase catalyzes oxidation of alcohol to produce aldehyde or ketone, with reduction of a cofactor such as NAD$^+$. In this case, most of the atoms in alcohol are conserved in aldehyde (or ketone), with the exception of the hydrogen atom and the electron, which are transferred to the cofactor. In addition, most of the atoms in NAD$^+$ are conserved in NADH. Reactant pairs for this reaction are alcohol−aldehyde (or ketone) and NAD$^+$−NADH. The H$^+$ ion associated with NAD$^+$ reduction is ignored. One compound may appear in two or more reactant pairs, if it is cleaved in the reaction. For example, a typical reaction of transferase

$$AB + C \Leftrightarrow A + BC$$

produces three reactant pairs: AB−A, AB−BC, and C−BC. Inorganic compounds may also be included in reactant pairs. For example, reactant pairs for a typical hydrolase reaction

$$AB + H_2O \Leftrightarrow AH + BOH$$

are AB−AH, AB−BOH, and H_2O−BOH. To which compound the oxygen atom that comes from water should be assigned is based on the knowledge of organic chemistry. In cases where the destination of the oxygen atom from water cannot be determined (e.g., hydration of glycosyl bond), the water molecule is assigned to the smaller product. For ligases, flux of one oxygen atom is ignored because of the difficulty to decide its destination. Reactant pairs for a typical ligase

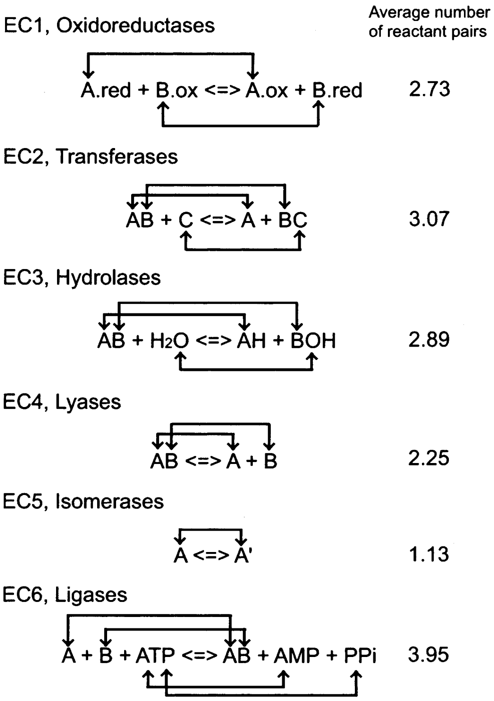

Figure 1. Extraction of reactant pairs from reactions of the six EC classes. Reactions in each class show unique topology of the reactant pairs, resulting in the average number of reactant pairs shown on the right.

$$A + B + ATP \Leftrightarrow AB + AMP + PPi$$

are A−AB, B−AB, ATP−AMP, and ATP−PPi.

Figure 1 shows a typical reaction and its description, with reactant pairs, in each of the six EC classes. The average number of reactant pairs per reaction is different for each class. Oxidoreductases (EC1) catalyze reactions consisting of main substrates and cofactors, which are decomposed typically into two pairs. However, there are many reactions consisting of cofactors with complex changes, for example, 1.14.x.x acting on paired donors with incorporation of molecular oxygen. This gives rise to an average of 2.73 pairs for each oxidoreductase reaction. The average number of pairs per reaction in other classes is almost comparable to the typical description of reactions shown in Figure 1.

Atom Typing (KEGG Atom Types). The COMPOUND section of the KEGG LIGAND database contains a collection of chemical compound structures, represented by the MDL-MOL file format, including those that appear in all 5227 reactions. A chemical structure is a graph object, where atoms and atomic bonds are represented as vertexes and edges, respectively, with the exception of hydrogen atoms, which are ignored. For the purpose of chemical structure comparison, we adopt atom typing whereby the same atomic species are distinguished based on the classification of functional groups. Thus, in our graph representation of chemical compounds, which we refer to as KEGG chemical function (KCF) representation, vertexes (atoms) are distinguished by 68 atom types called KEGG atom types (Table 1).[5] Generally, each KEGG atom type is labeled with three units. The first character represents the atomic species, such as C for carbon. The second numeral represents the electron environment. The third character represents the information on the substituted groups, and it is given serially to linear substructures (a−d) or to circular substructures (x−

[5] 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.

[6] Goto, S.; Nishioka, T.; Kanehisa, M. LIGAND: chemical database for enzyme reactions. *Bioinformatics* **1998**, *14*, 591−599.

Table 1. Atom Typing for Defining 68 KEGG Atoms

functional group	atom type	definition	functional group	atom type	definition
			Carbon (23 types)		
alkane	C1a	R−CH_3	alkyne	C3a	R≡C−H
	C1b	R−CH_2−R		C3b	R≡C−R
	C1c	R−CH(−R)−R	aldehyde	C4a	R−CH═O
	C1d	R−C$(−R)_2$−R	ketone	C5a	R−C(═O)−R
cyclic alkane	C1x	ring−CH_2−ring	cyclic ketone	C5x	ring−C(═O)−ring
	C1y	ring−CH(−R)−ring	carboxylic acid	C6a	R−C(═O)−OH
	C1z	ring−C$(−R)_2$−ring	carboxylic ester	C7a	R−C(═O)−O−R
alkene	C2a	R═CH_2		C7x	ring−C(═O)−O−ring
	C2b	R═CH−R	aromatic ring	C8x	ring−CH═ring
	C2c	R═C$(−R)_2$		C8y	ring−C(−R)═ring
cyclic alkene	C2x	ring−CH═ring	undefined C	C0	
	C2y	ring−C(−R)═ring			
			Nitrogen (16 types)		
amine	N1a	R−NH_2	cyclic imine	N2x	ring−N═ring
	N1b	R−NH−R		N2y	ring−N^+(−R)═ring
	N1c	R−N$(−R)_2$	cyan	N3a	R≡N
	N1d	R−$N^+$$(−R)_3$	aromatic ring	N4x	ring−NH−ring
cyclic amine	N1x	ring−NH−ring		N4y	ring−N(−R)−ring
	N1y	ring−N(−R)−ring		N5x	ring−N═ring
imine	N2a	R═NH		N5y	ring−N^+(−R)═ring
	N2b	R═N−R	undefined N	N0	
			Oxygen (18 types)		
hydroxy	O1a	R−OH	oxo	O3a	N═O
	O1b	N−OH		O3b	P═O
	O1c	P−OH		O3c	S═O
	O1d	S−OH	aldehyde	O4a	R−CH═O
ether	O2a	R−O−R	ketone	O5a	R−C(═O)−R
	O2b	P−O−R		O5x	ring−C(═O)−ring
	O2c	P−O−P	carboxylic acid	O6a	R−C(═O)−OH
	O2x	ring−O−ring	carboxylic ester	O7a	R−C(═O)−O−R
				O7x	ring−C(═O)−O−ring
			undefined O	O0	
			Sulfur (7 types)		
thiol	S1a	R−SH	disulfide	S3a	R−S−S−R
thioether	S2a	R−S−R		S3x	ring−S−S−ring
	S2x	ring−S−ring	sulfate	S4a	R−SO_3
			undefined S	S0	
			Phosphorus (2 types)		
attached to others	P1a	P−R	attached to oxygen	P1b	P−O
			Others (2 types)		
halogen	X	F, Cl, Br, and I	others	Z	

z). Hydrogen atoms and the bonds consisting of hydrogen atoms are not represented in the graph structure, but the numbers of attached hydrogen atoms are reflected in the vertexes that represent the hydrogenated atoms.

Chemical Structure Comparison Method. Each of the reactant pairs is subjected to the chemical structure comparison method that we developed before.[5] The approach is based on a graph theoretical method for finding common (isomorphic) subgraphs in two graphs of chemical compound structures. The actual algorithm involves finding the maximal cliques within the so-called association graph, which is defined by two initial graphs. Although our implementation of the algorithm is essentially the same as the traditional association graph methods,[7,8] we have incorporated some heuristics because mathematically strict solutions are sometimes found to be inadequate from the biochemical point of view. One is atom typing mentioned above, and the others include partial weighting of atom matches and threshold parameters for optimization in the clique finding processes. When two compounds are given, our method identifies the largest matched structure with the least number of boundary atoms between the matched and nonmatched structures and outputs the alignment of atoms (the list of matched or gapped atoms). For a reactant pair the boundary atom is considered as a reaction center, and the conversion patterns of KEGG atom types are recorded for the reaction center and its neighbors. The precise definition of the association graph and the detailed algorithm to align chemical structures are described in our original article.[5]

Results

Definition of R-, D-, and M-Atoms. Each reaction formula of known enzymatic reactions is decomposed into a set of reactant pairs, and chemical structure alignment of each reactant pair is performed to extract a conversion pattern. As a result, reactant pairs were divided into two groups, with and without the difference (nonmatched) structure, which is the structure that was not aligned. The assignment of RC numbers is based on the identification of the reaction center atom (R-atom), the difference structure atoms (D-atoms), and the matched structure atoms (M-atoms), as well as the conversion patterns of KEGG atom types. For the pair with the difference structure (Figure 2A), we define the R-atom as the atom that belongs to the matched structure and that is adjacent to the difference structure.

[7] Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

[8] Qu, D. L.; Fu, B.; Muraki, M.; Hayakawa, T. An encoding system for a group contribution method. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 443−447.

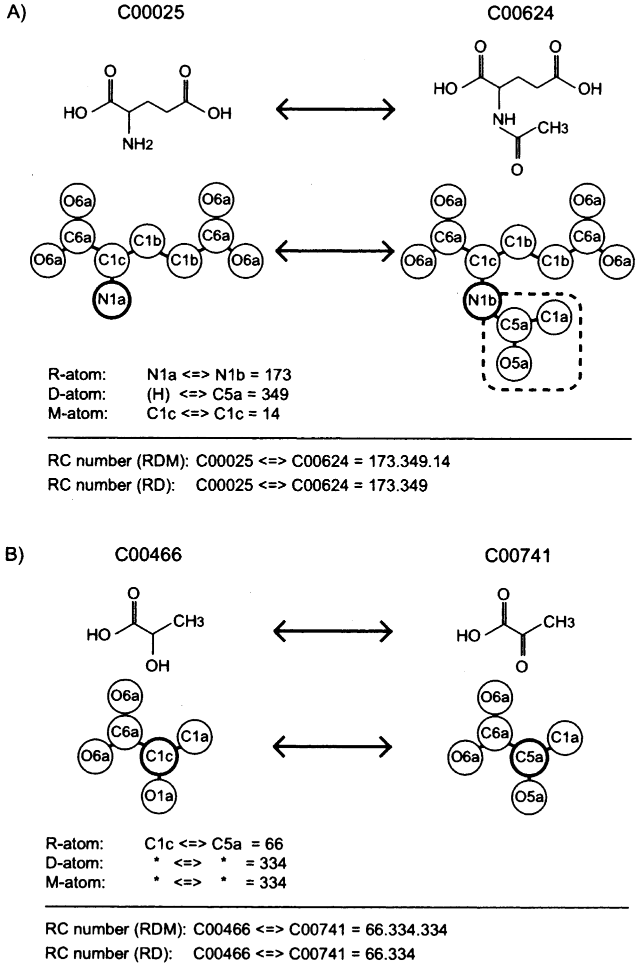
A)
C00025
C00624
HO
OH
NH2
NH
CH3
O
O6a
C6a
C1c
C1b
N1a
N1b
C5a
C1a
O5a
R-atom: N1a <=> N1b = 173
D-atom: (H) <=> C5a = 349
M-atom: C1c <=> C1c = 14
RC number (RDM): C00025 <=> C00624 = 173.349.14
RC number (RD): C00025 <=> C00624 = 173.349
B)
C00466
C00741
CH3
HO
OH
O
O1a
C5a
R-atom: C1c <=> C5a = 66
D-atom: * <=> * = 334
M-atom: * <=> * = 334
RC number (RDM): C00466 <=> C00741 = 66.334.334
RC number (RD): C00466 <=> C00741 = 66.334

Figure 2. Assignment of the RC number, which describes the conversion patterns of the KEGG atom types for the reaction center atom (R-atom), the difference structure atom (D-atom), and the matched structure atom (M-atom). The definition of the R-, D-, and M-atoms are somewhat differentt for the cases of (A) a partial match with the difference structure (surrounded by dashed line) and (B) a complete match without the difference structure. See text for more details.

The atoms that are adjacent to the reaction center and belong to the difference and matched structures are referred to as D- and M-atoms, respectively. For example, in the reactant pair C00025−C00624 (Figure 2A), the R-, D-, and M-atoms are identified as shown. Furthermore, the conversion patterns of KEGG atom types are N1a ⇔ N1b, (H) ⇔ C5a, and C1c ⇔ C1c, respectively, for the R-, D-, and M-atoms, where (H) means the hydrogen atom was ignored in the KCF representation. The numerical representation of these conversion patterns, 173, 349, and 14, are used to define the RC number, in the form of either RDM or RD. Note that there are cases where multiple M-atoms or multiple D-atoms can be defined for the single R-atom.

For the pair without the difference structure (Figure 2B), the R-, D-, and M-atoms are defined in a different way. Because addition and elimination of hydrogen atoms are not described as changes in the KCF structure topology, but represented as changes of KEGG atom types, the R-atom is defined as the atom belonging to the matched structure, with a change of atom typing, and adjacent to atoms without any changes of atom typing. The D- and M-atoms are represented as asterisks (*⇔*) meaning only the changes of attached hydrogen atoms and electron environments.

Assignment of the RC (Reaction Classification) Numbers. The conversion patterns of 68 KEGG atom types for the R-, D-, and M-atoms are collected and represented by numeral codes (Supporting Information 1), which are numbered sequentially from 1 to 1422 for all the patterns derived from our dataset. For example, the conversion from N1a to N1b is given 173 as shown in Figure 2A. (The total number of observed patterns for the R-, D-, and M-atoms is summarized in the Supporting Information 1.) The reaction classification (RC) number is a combination of these numerals separated by periods, representing the conversion patterns of the R-, D-, and M-atoms. Note that the RC number does not represent any hierarchy, as in the case of the EC number. It only represents the reaction classification with varying details, the reaction center only with the first numeral (R), considering the difference region with the first two numerals (RD), further considering the matched region with the three numerals (RDM).

Our dataset consists of 3254 EC numbers. The first three numerals of the EC number indicate the hierarchical classification of reactions, one of the six classes, the subclass, and the sub-subclass, while the last numeral is a serial number for substrate specificity. To compare with our reaction classification, it is thus appropriate to consider up to the third numeral for the sub-subclass. Then our dataset contains 214 different types of reactions (sub-subclasses) according to the EC classification. In our RC classification, the numbers of R, RD, and RDM entries obtained are 225, 635, and 1018, respectively, suggesting that the RC system provides more detailed classification than the EC system.

Representation of Reaction Formula. A reaction formula corresponding to an EC number is decomposed into a set of reactant pairs, which are then converted to a set of RC numbers. During this process, the information about how the reactant pairs are defined from the original reaction formula is not preserved. Figure 3 illustrates an example of a transferase reaction. In this reaction, the acetyl group of compound C00024 is transferred to the amine of compound C00025, producing C00624 and C00010 (Figure 3A). Reactant pairs for this reaction are C00025−C00624, C00024−C00624, and C00024−C00010 (Figure 3B). These pairs are converted to the RC numbers 173.349.14, 61.718.8, and 298.349.8, respectively. Here, some of the RC numbers relate to each other because the corresponding reactant pairs have a compound in common. Thus, to fully represent the original reaction formula, we need to consider a graph object whose vertexes and edges are compounds and RC numbers, respectively (Figure 3C), which we refer to as the "RC combination".

Correlation between the RC Numbers and the EC Numbers. We then examined the degree of correlation between the RC numbers and the EC numbers, in the hopes of utilizing our scheme for automatic assignment of EC numbers, as well as for extracting reaction characteristics. We found that most full EC numbers give rise to a single RC combination (Table 2). In contrast, EC sub-subclasses include a large variation in the RC combinations, which suggests that the RC system tends to classify reactions more strictly than the EC system in the viewpoint of the reaction classification. The average numbers of RC combinations (Figure 3C) per sub-subclass are 8.75, 33.5, 15.6, 24.0, 9.12, and 8.09, respectively, for the EC classes of 1 to 6 (Table 2). It is remarkable that the sub-subclasses of

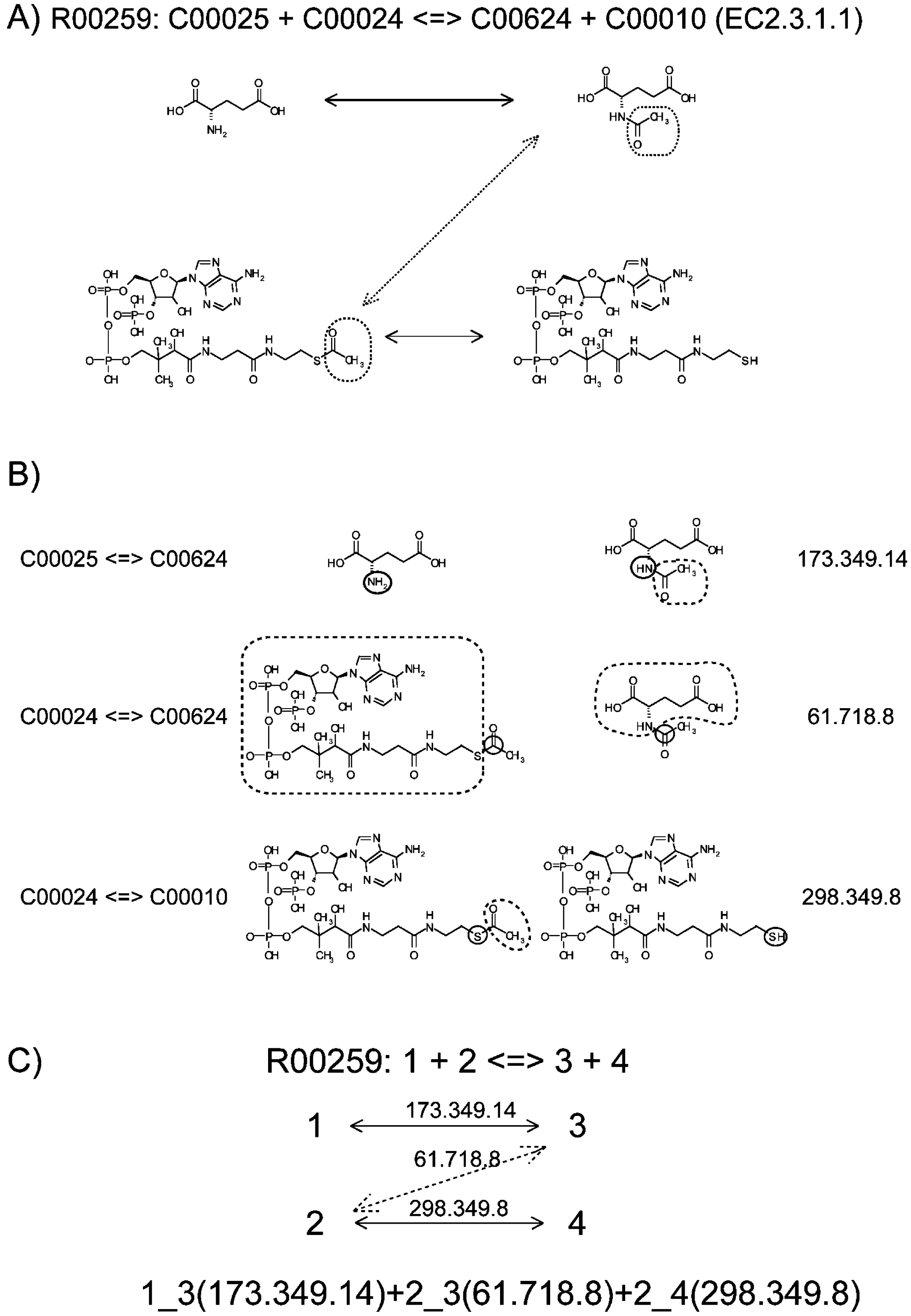

Figure 3. RC representation of an enzymatic reaction. (A) The reaction formula for R00259, a transferase reaction, is decomposed into three reactant pairs as indicated by arrows. Atom groups surrounded by dotted lines indicate transferred groups. (B) Each of the three pairs is subjected to the chemical structure comparison to identify reaction centers and difference structures marked by solid and dashed lines, respectively. Each pair is thus assigned an RC number. (C) The RC combination of the reaction is a graph representation where nodes are represented by numerals (without any chemical compound information), and edges are arrows associated with the RC numbers.

transferases (EC2) include a larger variety of reactions than all other classes, because transferase reactions have three different characteristics: the transferred group, the donor and the acceptor group of the transferred group. (The correspondence between each EC sub-subclass and one or more RC combinations is shown in Supporting Information 2.)

Usefulness of our RC system is also examined by the jackknife cross-validation test. In advance, we took away reactions unsuitable for this test. We did not use: (i) reactions without the EC numbers assigned up to the sub-subclasses (third numeral), (ii) reactions belonging to the subclass or sub-subclass of “97” or “99” meaning miscellaneous substrates or reactions, (iii) reactions containing compounds without KCF structures such as glycans or proteins, and (iv) reactions belonging to the sub-subclasses with single members which are obviously unsuitable for the jackknife (leave-one-out) procedure. Of the 5227 reactions, the total number of reactions that satisfy these conditions was 4570.

To analyze the correlation between the EC and the RC classification systems, we examined six cases for the description

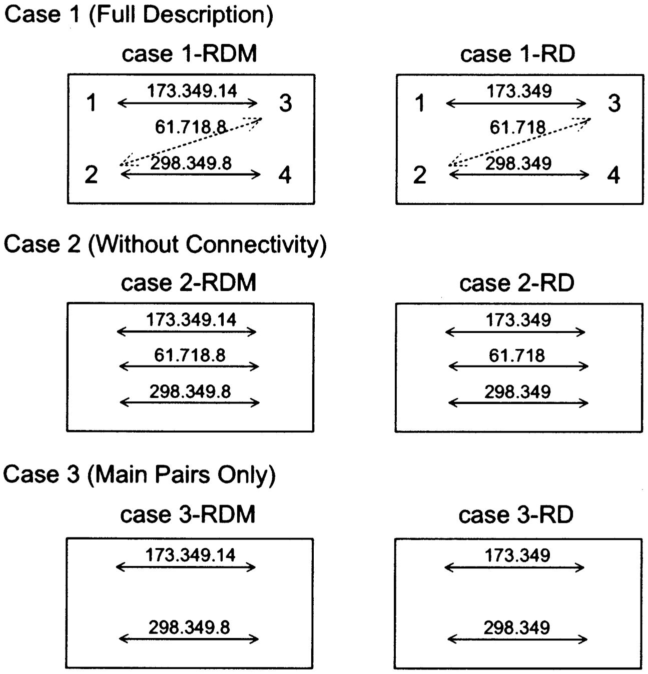

Figure 4. Six cases of RC description of enzymatic reactions for the jackknife cross-validation test. The distinction is made for the level of knowledge about the reactant pairs (full description, without connectivity, or main pairs only) and the level of details about the conversion patterns (RDM or RD).

Table 2. Correlation between the RC Numbers and the EC Numbers[a]

	number of		number of RC combinations for	
EC class	sub-subclasses	full EC numbers	sub-subclass	full EC number
1. oxidoreductases	78	901	8.75	1.46
2. transferases	26	1024	33.5	1.50
3. hydrolases	33	564	15.6	1.58
4. lyases	12	300	24.0	1.53
5. isomerases	12	125	9.12	1.26
6. ligases	10	130	8.09	1.27
total	171	3044	14.9	1.49

[a] The number of RC combinations (Figure 3C) is calculated for each class, without miscellaneous subclasses and sub-subclasses.

of reactions with the RC numbers as shown in Figure 4. First, we distinguish three cases depending on how much information is given about the reaction, full description when both the RC numbers and their connectivity patterns are given (case 1), without connectivity when the RC numbers are given (case 2), and main pairs only when a partial set of the RC numbers is given (case 3). Second, we distinguish the detail of the RC numbers in the forms of RDM and RD (Figure 2). Note that case 1-RDM is what we call the RC combination representing complete description of reaction characteristics (Figure 3C) and that the other five cases are incomplete descriptions for dealing with reactions whose whole characteristics are not known. In practice, the connection pattern (case 1) can usually be assumed from the given set of compounds (case 2) by linking the same compounds. However, this does not always work because the same compound may appear on both sides of the reaction and because reaction stoichiometry may have to be taken into account.

Table 3. Result of the Jackknife Cross-Validation Test

		accuracy of the prediction (%)			
condition	coverage (%)	class	subclass	sub-subclass	full EC number
1-RDM	62.4	98.5	93.2	89.1	19.4
1-RD	68.5	98.2	92.8	88.2	15.8
2-RDM	69.4	92.1	84.9	80.3	17.6
2-RD	75.5	92.1	85.0	80.0	14.3
3-RDM	78.2	84.0	77.1	67.3	13.0
3-RD	84.2	83.5	76.3	66.0	10.1

Each reaction in the dataset of 4570 reactions is taken as a query, represented in the six cases, and compared to the remaining 4569 reactions. When the query reaction has hits, namely, reactions sharing the same RC cases are found, the EC numbers are compared between those of the hits and the real EC number of the query. The rate of correlation between the RC system and the EC system, which we call the accuracy of EC number prediction, is defined by simply summing up all correct and incorrect hits in each query. The coverage of EC number prediction is the rate of queries that have hits. As shown in Table 3, the accuracy is computed for all hierarchical levels of the EC numbers with the six different cases. The strictest condition (case 1-RDM) has an accuracy of 89.1% with a coverage of 62.4% for the EC sub-subclasses. The loosest condition (3-RD) has an accuracy of 66.0% with a coverage of 84.2% in the same category. In practice, the queries that have no matching reactions under strict conditions may be consequently examined with looser conditions.

There are some inconsistencies between the results obtained by the RC system and the EC system. In the RC classification system, two aspects remain to be improved for more accurate assignment. One is improvement of the molecular alignment method, which would be better if some aspects of the whole reaction formula were considered rather than separate alignments of individual reactant pairs alone. The other is the description of the R-, D-, and M-atoms, where more specific knowledge of organic chemistry may be included to define these atoms. As shown below, the EC system may also need improvements in terms of consistency.

Examples of Single EC Sub-Subclasses Corresponding to Multiple RC Combinations. The EC numbers and the RC numbers focus on somewhat different aspects of reactions. There are cases where two reactions belonging to the same EC sub-subclass have different RC combinations (Table 2) and where two reactions belonging to different EC sub-subclasses have the same RC combination. Figure 5 shows some examples of the former cases.

Both R02138 and R04956 belong to 2.3.1.x, but their reactant pairs differ significantly (Figure 5A). Although the two reactions R02138 (C00154 + C00381 ⇔ C00010 + C02536) and R04956 (C05749 + C00006 ⇔ C05748 + C00005) are difficult to distinguish at the level of reaction formulas, the differences are easy to spot using reactant pair descriptions. There are three pairs for R02138 (C00154−C00010, C00154−C02536, and C00381−C02536) and two pairs for R04956 (C05749−C05748 and C00006−C00005). This is caused by the fact that the EC number for a multistep reaction is assigned to its elementary steps.

When we consider not only the number of reactant pairs but also the description of the R-atoms, we can distinguish the

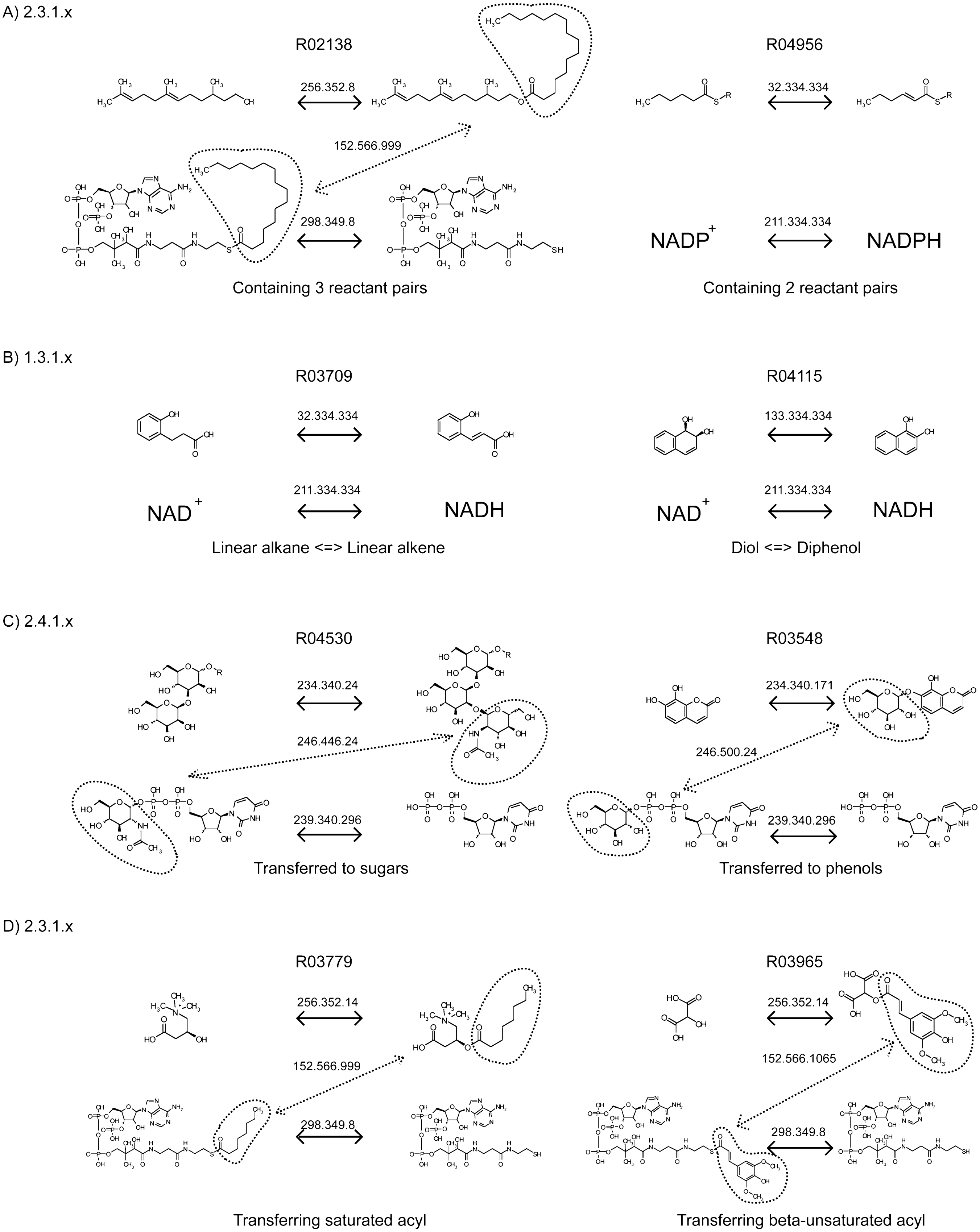

Figure 5. Examples of single EC sub-subclasses corresponding to multiple RC combinations. Many of the EC sub-subclasses include a variety of RC combinations from the viewpoint of (A) the topology of the reactant pairs, (B) R-atoms, (C) D-atoms, and (D) M-atoms. See text for more details.

difference among reactions more precisely (Figure 5B). For example, 1.3.1.x represents enzymes that oxidize saturated carbon−carbon bonds (C−C) to produce unsaturated bonds (C=C) using NAD(P)$^+$. When considering R-atoms, we found that there are several different types of reactant pairs in this sub-subclass including alkane−alkene (R03709) and diol−

diphenol (R04115). It is a matter of choice whether to distinguish these two types of reactions, but generation and breakage of aromatic rings are important events, especially for secondary metabolism such as involving polyphenols.

More detailed classification can be obtained by considering the D-atoms (Figure 5C). For example, 2.4.1.x represents transferases that transfer hexose residues. R04530 and R03548 are classified in the same group when considering only the R-atoms but as distinct groups when considering the D-atoms, which reflects the destination of the transferred groups: to sugars (R04530) and to phenols (R03548). In the EC system, the sub-subclasses of transferases are usually classified based on the transferred groups, but some transferases are classified in more detail. For example, transferases acting on phosphorus-containing groups, such as phosphotransferases and nucleotidyltransferases, are further classified based on the destination of the transferred group or substrate specificity. In the RC system, we can deal with all of the transferases in a unified manner: transferred groups and the donor and the acceptor of the transferred group, without considering substrate specificity or other characteristics.

Considering the M-atoms provides more specified classification (Figure 5D). For example, 2.3.1.x represents transferases that transfer acyl groups other than amino-acyl groups. Even with considering the R- and D-atoms, R03779 and R03965 are classified into the same group. However, when M-atoms are taken into account, they are distinguished based on the variety of atoms neighboring the R-atoms: β-saturated acyl for R03779 and β-unsaturated acyl for R03965.

Examples of Single RC Combinations Corresponding to Multiple EC Sub-Subclasses. Although the RC system tends to classify enzymes in a more detailed way than the EC system (Table 2), even the strictest description (1-RDM in Figure 4), that is, the RC combination, is found to be shared by different EC sub-subclasses. The average numbers of sub-subclasses per RC number (RDM in Figure 2) and per RC combination (1-RDM in Figure 4) are 2.00 and 1.08, respectively. Figure 6 shows some examples of reactions that share the same RC combination but have different EC sub-subclasses. (Supporting Information 3 shows the correspondence between each RC combination and one or more EC sub-subclasses.)

In some cases, enzymes with the same type of reactions are classified in different EC subclasses or sub-subclasses (Figure 6A). 1.2.1.x represents oxidoreductases acting on carbonyl groups (aldehydes or ketones) with NAD(P)$^+$ as acceptor. Typical 1.2.1.x enzymes produce carboxylic acid. However, some enzymes in this sub-subclass represent alcohol ⇔ ketone conversions (R02260, for example), which are the same as typical 1.1.1.x reactions.

Some of the multistep reactions give rise to a situation where the same types of reactions belong to different EC sub-subclasses (Figure 6B). One example is observed in ligases (6.x.x.x), which typically catalyze the conjugation of two substrates with hydrolysis of a nucleoside triphosphate such as ATP. This ligase reaction consists of two types of reactions, as for AMP-forming acetate-CoA ligase (6.2.1.1) catalyzing two reactions: R00316 and R00236. On the other hand, each of these reactions is just the same as a transferase reaction: R00315 (2.7.2.1) and R00230 (2.3.1.8), total of which comprises the typical ligase reaction. Apparently, the reason for these different EC assignments is based on the consecutiveness of the reactions. The intermediate compound in the former case (6.2.1.1) is kept combined with the enzyme, while it is released in the latter case (2.7.2.1 and 2.3.1.8).

Other types of common properties between certain EC sub-subclasses are concerned with substrate specificity (Figure 6C). The subclass 2.4.x.x represents glycosyltransferases, and the third numeral distinguishes between hexosyltransferases (2.4.1.x) and pentosyltransferases (2.4.2.x). However, these two categories share the same type of reactions, for example, R04530 for 2.4.1.x and R03268 for 2.4.2.x. Although substrate specificity is supposed to be represented by the fourth numeral, rather than the third numeral, in the EC system, there are exceptions or different views. The RC system enables one to distinguish the sort of chemical bonds involved in reactions but does not distinguish between substrates that are hexoses and pentoses.

Discussion

Use of the RC System. We demonstrated that the enzymatic reactions in the known dataset can be distributed to correct EC numbers by the computational method using the RC system, which enables us to assign EC numbers faster and more accurately than the manual assignment has done before. Reactions can be assigned using our RC system even if full characteristics of enzymes or enzymatic reactions are not known. This is a great advantage especially for the classification of enzymes that are hard to characterize experimentally or that are known only from the pathways of main compounds.

An advantage of our RC system is the novel description of enzymatic reactions, a graph object representation consisting of compounds as vertexes and RC numbers as edges. The reaction formula associated with the EC number, such as

$$\text{Substrate\_1} + \text{Substrate\_2} \Leftrightarrow \text{Product\_1} + \text{Product\_2}$$

is not enough to represent flow of atoms,[9] especially in the case where the full characteristics are not described. In contrast, our method is based on representing an enzymatic reaction as a combination of elementary reaction steps involving reactant pairs, such as

$$\text{(Substrate\_1, Product\_1), (Substrate\_2, Product\_1), and (Substrate\_2, Product\_2)}$$

which can be utilized in flux analyses and other types of metabolome analyses. The dataset of reactant pairs in all known enzymatic reactions that has been created in the present study, containing the information about the chemical structure alignments and the RC numbers, is made available as part of the KEGG database (http://www.genome.jp/kegg/).

Another possible use of our RC method is to resolve missing enzymes in the metabolic reconstruction. When the genome is completely sequenced for an organism, enzyme genes with predicted EC numbers can be mapped onto known metabolic pathway diagrams such as those provided by KEGG, and the organisms's metabolic capability can be deduced from a set of pathways that are completely reconstructed. Sometimes, a certain

[9] Arita, M. In silico atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. *Genome Res.* **2003**, *13*, 2455−2466.

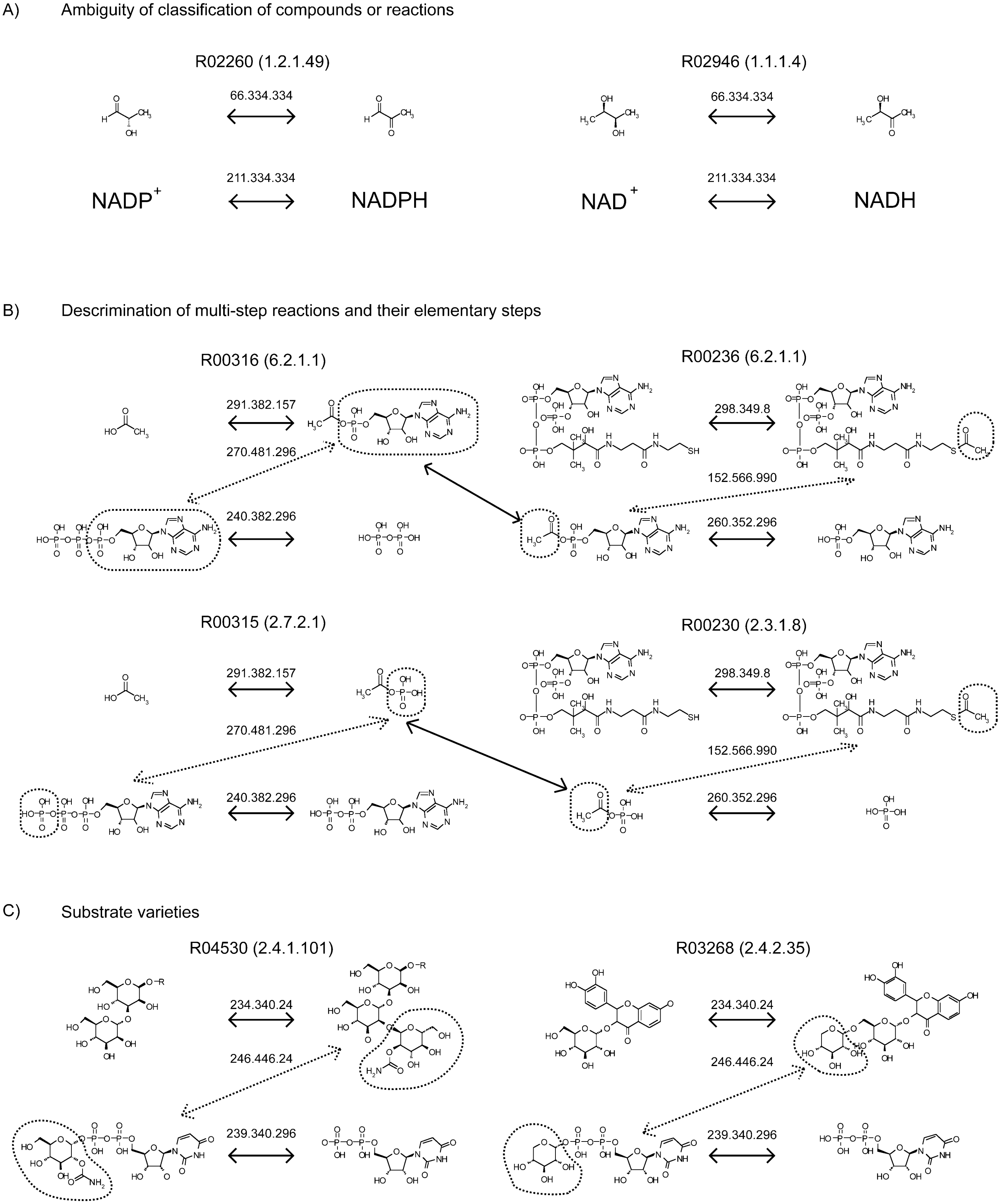

Figure 6. Examples of single RC combinations corresponding to multiple EC sub-subclasses. The RC system detects common characteristics among different EC sub-subclasses. This can be explained by: (A) ambiguity of classification of compounds or reactions, (B) discrimination of multistep reactions and their elementary steps, and (C) substrate varieties. See text for more details.

pathway is almost complete but contains a few gaps of missing enzymes, in which case our RC method can be applied to substrate−product pairs to suggest additional EC numbers, which can be re-examined in the genomic sequence.

Prediction of EC Numbers for Unassigned Reactions. To evaluate the ability of our method to predict EC numbers, we applied the procedure described in the cross-validation test (Table 3) to those enzymes that are not yet fully categorized.

Table 4. Examples of Reactions Belonging to the Same OC and RC Combination but to Different EC Sub-subclasses

reactions	common properties of reactions	reactions of different EC
2.3.1.-(R01701) and 1.2.4.2 (R01700)	2-Oxo-acid is decarboxylated and transferred onto a thiol group.	These reactions are followed by oxidative cleave to produce carboxylate, the total of which are regarded as oxidative decarboxylation of 2-oxo-acid.
1.2.1.3 (R00710) and 1.5.1.12 (R00245)	Aldehydes are oxidized to produce carboxylate with reducing NAD$^+$.	Main substrates of R00245 includes CH−NH group.
4.2.1.-(R04417) and 4.3.1.17 (R00590)	A water molecule is dissociated from hydroxy group to produce unsaturated carbon−carbon bond.	Main substrate of R00590 includes amino group.
1.5.1.20 (R01224) and 1.7.99.5 (R01223)	Amine is dehydrogenized to produce cyclic amine with reducing NAD(P)$^+$.	Both substrate and cofactor in R01223 are regarded as “other” compounds.
1.13.11.43 (R00043) and 1.14.99.36 (R00032)	Molecular oxygen is incorporated into unsaturated carbon−carbon bond, which causes separation of substrate to produce two aldehydes.	R00032 is classified into “miscellaneous” reactions.
3.4.13.3 (R01166) and 3.5.1.18 (R02734)	Hydration of peptide or amide bond.	Peptide and amide bond are chemically same, but they belong to distinct subclasses in EC, and classified into different ways in their sub-subclasses.

There are 296 reactions belonging to the EC categories described as “99”, which means “other” sub-subclasses. Each of these reactions is represented by the RC combination, or a most strict RC description of Figure 4 when it is not available, and compared against our dataset, which did not include the “99” categories. Out of 296 reactions, we assigned putative EC numbers to 213 reactions, including those cases where EC numbers could not be determined uniquely. We found that 66 reactions should be classified in well-defined EC sub-subclasses and that 136 reactions are also relatively well-defined although they lack information about cofactors. Results of the assignment can be found in Supporting Information 4.

In the KEGG REACTION database there are other 480 unassigned reactions, mostly taken from the KEGG PATHWAY database. Among them we found that 136 reactions have identical RC combinations to known reactions, and the EC sub-subclasses could be assigned accordingly (Supporting Information 4). There were also cases where EC numbers could not be assigned properly, indicating new types of reactions that are not present in the current EC system or reflecting ambiguity of the EC system such as that mentioned in Figure 6A. Even though there are no matching sub-subclasses with the same RC combinations, similar sub-subclasses may be found if a proper measure of similarity can be defined for the RC combinations. Thus, our RC method may be used to suggest the necessity of defining new EC sub-subclasses for undefined types of reactions.

Correspondence to Protein Sequences. We have shown that the EC system and the RC system represent different classification schemes with different points of view. Then an obvious question is which is better correlated to protein classifications based on sequence information. We used KEGG ortholog clusters (OCs), which are computationally identified clusters of orthologous genes in the KEGG GENES database containing all the genes in the completely sequenced genomes. The OCs were obtained by graph analysis of clique searching in the KEGG SSDB database containing sequence similarity scores among all those genes. Thus, the total of 585 184 genes in 170 genomes was decomposed into 36 165 OCs excluding singletons. Among these, the OCs that include genes for enzymes was 3413, which was used for our analysis.

It was found that almost 90% of the OCs containing enzymes were related to a single EC number, but less than 20% of the EC sub-subclasses were included in single OCs meaning that enzymes within the same sub-subclass do not necessarily share sequence similarity. This suggests that protein families (ortholog clusters) mostly represent substrate specificity rather than reaction specificity.[10] However, since most of the putative enzyme genes unveiled in the genome sequencing projects are annotated based on sequence similarity, it is not surprising at all to observe a good correlation between ortholog clusters and full EC numbers. Our RC system is intended for analysis of reaction specificity, excluding substrate specificity. Using this we could find cases where enzymes with similar sequences are assigned to different EC categories but have the same type of reaction. Table 4 shows such examples; enzymes belonging to the same OC have exactly the same RC combination but belong to different EC sub-subclasses. There are two possible reasons for this. One is due to ambiguity in the assignment of the EC numbers, and the other is caused by the distinction between multistep reactions and their elementary steps (see Figure 6B).

There are some other interesting cases of partially identical RC combinations, where enzymes share similar reactivity and similar sequences but do not belong to the same EC sub-subclass. Table 5 shows such additional examples, which reflect enzymes catalyzing the same type of reaction in an inter- or intramolecular way, working on the same type of chemical bond for transferases or hydrolases, and concerning multistep reactions. More detailed analysis of these examples will help to understand the diversity of enzymatic reactions and the diversity of enzyme genes, especially in the secondary metabolism.

There have been works on analyzing the correlation between the EC system and protein structures. Enzymes in different EC categories share similar chemical capabilities from the viewpoint of 3D structures,[11] which also implies the necessity of developing a new classification of enzymatic function based solely on reaction specificity. There are also studies that investigate the correlation between protein sequences (or structures) and functions[12−16] using the hierarchical EC classification system or reconstruction of metabolic pathways using gene contents

[10] Cai, C. Z.; Han, L. Y.; Ji, Z. L.; Chen, Y. Z. Enzyme family classification by support vector machines. *Proteins* **2004**, *55*, 66−76.

[11] Babbitt, P. C. Definitions of enzyme function for the structural genomics era. *Curr. Opin. Chem. Biol.* **2003**, *7*, 230−237.

[12] Shakhnovich, B. E.; Max Harvey, J. Quantifying structure−function uncertainty: a graph theoretical exploration into the origins and limitations of protein annotation. *J. Mol. Biol.* **2004**, *337*, 933−949.

[13] Devos, D.; Valencia, A. Practical limits of function prediction. *Proteins* **2000**, *41*, 98−107.

Table 5. Examples of Reactions in the Same OC and Including the Same RC Number but with Different EC Sub-subclasses

reactions	common properties of reactions	reasons of different EC
3.1.3.13 (R01516) and 5.4.2.1 (R01662)	Phosphoric monoester is cleaved.	In inter- (EC3) and intra- (EC5) molecular way.
3.1.3.18 (R01334) and 5.4.2.6 (R02728)		
2.7.1.105 (R00757) and 3.1.3.46 (R00763)	Phosphoric ester is cleaved.	Transferred to water (EC3) and other compounds (EC2).
2.7.7.43 (R01117) and 3.1.3.29 (R01805)		
2.7.6.5 (R00429) and 3.1.7.2 (R00336)		
2.7.7.-(R05222) and 3.6.1.-(R04549)		
2.7.1.60 (R02705) and 5.1.3.14 (R00414)		
2.3.1.65 (R03718) and 3.1.2.2 (R01274)	Acyl-thioester is cleaved.	Transferred to water (EC3) and other compounds (EC2).
2.1.2.2 (R04325) and 3.5.1.10 (R00944)	N-formyl group is cleaved.	Transferred to water (EC3) and other compounds (EC2).
2.1.2.2 (R04325) and 6.3.4.13 (R04144)	N-acyl group is synthesized.	With (EC6) and without (EC2) hydrolation of ATP.
4.1.1.3 (R00217) and 6.4.1.1 (R00344)	Carboxylate is decarboxylated.	With (EC6) and without (EC4) hydrolation of ATP.
4.1.1.41 (R00923) and 6.4.1.3 (R01859)		
6.3.4.14 (R04385) and 6.4.1.4 (R04138)	Carbon dioxide is incorporated with hydrolation of ATP.	Carboxylate is formed on nitrogen (EC6.3) or carbon (EC6.4).
2.6.1.19 (R00908) and 5.4.3.8 (R02272)	Amino and oxo group are exchanged.	In inter- (EC2) or intra- (EC5) molecular way.
2.6.1.16 (R00768) and 3.5.99.6 (R00765)	Amino and oxo group are exchanged.	Amino/oxo groups are related to water/ammonia (EC3) or other compounds (EC2).
2.6.1.5 (R00694) and 4.4.1.8 (R02408)	Amino and oxo group are exchanged.	With (EC4) and without (EC2) being followed by cleavage.
4.1.3.27 (R00986) and 6.3.5.2 (R01231)	Amino and oxo group are exchanged.	With (EC6) and without (EC4) hydrolation of ATP.
2.4.2.14 (R01072) and 2.6.1.16 (R00768)	Amino and hydroxy group are exchanged.	With (EC2.4) and without (EC2.6) transferration of sugars.

and their functions.[17-19] However, these studies did not consider the situation that some EC sub-subclasses include more diverse reactions and that even different EC classes (the first figures) sometimes share reaction properties, as described above.

Possible Extensions of the EC Numbers. In the EC system the criteria of classifying subclasses and sub-subclasses are somewhat dependent on different categories. Well-known enzymes, such as phosphotransferases, nucleotidyltransferases, nucleases, and peptidases, tend to be classified in more detail than other enzymes. In contrast, those enzymes that have recently been uncovered in more detail, such as glycosyltransferases,[20] are not given detailed classification criteria in the current EC system. The criteria of distinguishing substrate specificity are based on published articles on enzymes; thus, the last figures of the EC numbers sometimes represent too general descriptions or too detailed distinctions. Close to one-half of the current EC numbers are not represented in the protein sequence databases, despite the fact that most major genomes are already completely sequenced, because enzyme genes are usually linked to those EC numbers with broader substrate specificity. Despite these irregularities, the EC system is so well established and widely used that we should try to improve it by computational methods, such as by our RC system.

Our present method is not suitable to assign the fourth figure of the EC numbers representing substrate specificity, as indicated by considerably low prediction rates in Table 3. Because the current EC system simply assigns serial numbers for substrate specificity, it would be difficult to computationally predict the last figure unless there is sufficient similarity of compound structures. Although the current EC numbers may already contain most enzymatic reactions (EC sub-subclasses), they may cover only a fraction of substrates that are present in nature. Thus, as an extension of the current work we are developing a classification scheme for substrates based solely on chemical structures, i.e., without using protein sequence information. This will make it possible to reclassify the last figures of the current EC numbers, to assign new reactions and substrates in more detail, and hopefully to better understand relations between

[14] Orengo, C. A.; Pearl, F. M.; Bray, J. E.; Todd, A. E.; Martin, A. C.; Lo Conte, L.; Thornton, J. M. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.* **1999**, *27*, 275−279.

[15] Todd, A. E.; Orengo, C. A.; Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol*. **2001**, *307*, 1113−1143.

[16] Hegyi, H.; Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol*. **1999**, *288*, 147−164.

[17] Ma, H.; Zeng, A. P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **2003**, *19*, 270−277.

[18] Goesmann, A.; Haubrock, M.; Meyer, F.; Kalinowski, J.; Giegerich, R. PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics* **2002**, *18*, 124−129.

[19] 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.

[20] Coutinho, P. M.; Deleury, E.; Davies, G. J.; Henrissat, B. An evolving hierarchical family classification for glycosyltransferases. *J. Mol. Biol*. **2003**, *328*, 307−317.

genomic diversity and chemical diversity of the metabolic pathways.

Acknowledgment. We thank Drs. Keith Tipton, Sinead Boyce, and Andrew McDonald for helpful comments and critical reading of the manuscript. Thanks are also due to Tomoko Komeno and Rumiko Yamamoto for collecting and curating reaction data and to Koichiro Tonomura for managing the LIGAND database system. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Supporting Information Available: The list of numerical codes for conversion patterns of KEGG atom types, the correspondence between each EC sub-subclass and different RC descriptions, and the correspondence between each RC description and different EC sub-subclasses. This material is available free of charge via the Internet at http://pubs.acs.org.

JA0466457