J|A|C|S
ARTICLES
Published on Web 09/06/2003

Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways

Masahiro Hattori, Yasushi Okuno, Susumu Goto, and Minoru Kanehisa*

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

Received May 9, 2003; E-mail: kanehisa@kuicr.kyoto-u.ac.jp

Abstract: Cellular functions result from intricate networks of molecular interactions, which involve not only proteins and nucleic acids but also small chemical compounds. Here we present an efficient algorithm for comparing two chemical structures of compounds, where the chemical structure is treated as a graph consisting of atoms as nodes and covalent bonds as edges. On the basis of the concept of functional groups, 68 atom types (node types) are defined for carbon, nitrogen, oxygen, and other atomic species with different environments, which has enabled detection of biochemically meaningful features. Maximal common subgraphs of two graphs can be found by searching for maximal cliques in the association graph, and we have introduced heuristics to accelerate the clique finding and to detect optimal local matches (simply connected common subgraphs). Our procedure was applied to the comparison and clustering of 9383 compounds, mostly metabolic compounds, in the KEGG/LIGAND database. The largest clusters of similar compounds were related to carbohydrates, and the clusters corresponded well to the categorization of pathways as represented by the KEGG pathway map numbers. When each pathway map was examined in more detail, finer clusters could be identified corresponding to subpathways or pathway modules containing continuous sets of reaction steps. Furthermore, it was found that the pathway modules identified by similar compound structures sometimes overlap with the pathway modules identified by genomic contexts, namely, by operon structures of enzyme genes.

Introduction

Whole genome sequencing has uncovered gene repertoires for more than a hundred organisms, but it has also clarified the needs for analyzing cellular functions as behaviors of a complex system rather than simply as a collected body of molecular functions.[1] The system of our interest is an interaction network of proteins, chemical compounds, and other components, which are also interacting with dynamic environments. Thus, it is an important problem to develop computational methods for analyzing large interaction networks and to understand systemic aspects of biology.[2,3] Coupled with computational approaches, significant efforts are undertaken for developing high-throughput experimental technologies and producing large-scale data in transcriptome,[4] proteome,[5] and metabolome analyses.[6] Furthermore, knowledge on chemical compounds, reactions, and pathways in cellular processes is accumulated in several biological databases, notably in KEGG.[7,8] In another attempt the categorization of genes in the context of higher-level functions is studied in Gene Ontology.[9,10] These database resources represent our current, probably very limited, knowledge on molecular interaction networks in living cells and organisms, but they can be used as reference knowledge from which we should be able to explore unknown networks by systematic analyses on large-scale experimental data.

The sequence-based methods for comparing genes and proteins are well-established, and we already have a picture on the "gene universe" in terms of the number of ortholog groups as reported, for example, in COG.[11] Similarly, established methods for three-dimensional (3D) structure comparisons provide a picture on the "protein universe" in terms of the

(1) Kanehisa, M.; Bork, P. Bioinformatics in the post-sequence era. *Nat. Genet.* **2003**, *33*, 305−310.
(2) Eisenberg, D.; Marcotte, E. M.; Xenarios, I.; Yeates, T. O. Protein function in the post-genomic era. *Nature* **2000**, *405*, 823−826.
(3) Kanehisa, M. Prediction of higher order functional networks from genomic data. *Pharmacogenomics* **2001**, *2*, 373−385.
(4) Velculescu, V. E.; Zhang, L.; Zhou, W.; Vogelstein, J.; Basrai, M. A.; Bassett, D. E., Jr.; Hieter, P.; Vogelstein, B.; Kinzler, K. W. Characterization of the yeast transcriptome. *Cell* **1997**, *88*, 243−251.
(5) Wilkins, M. R.; Sanchez, J. C.; Gooley, A. A.; Appel, R. D.; Humphery-Smith, I.; Hochstrasser, D. F.; Williams, K. L. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.* **1996**, *13*, 19−50.
(6) Tweeddale, H.; Notley-McRobb, L.; Ferenci, T. Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. *J. Bacteriol.* **1998**, *180*, 5109−5116.
(7) Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **1997**, *13*, 375−376.
(8) Kanehisa, M.; Goto, S.; Kawashima, S.; Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **2002**, *30*, 42−46.
(9) Schulze-Kremer, S. Ontologies for molecular biology. *Pac. Symp. Biocomput.* **1998**, *3*, 693−704.
(10) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25−29.
(11) Tatusov, R. L.; Koonin, E. V.; Lipman, D. J. A genomic perspective on protein families. *Science* **1997**, *278*, 631−637.

number of unique folds in SCOP[12,13] or CATH.[14] In contrast, we have little knowledge on the "chemical universe" consisting of chemical compounds and reactions in biological processes. In fact, there have been few analyses on comparison and classification of chemical compounds from a biological viewpoint, despite the fact that small chemical compounds are as important as biological macromolecules of proteins and nucleic acids in understanding molecular interaction networks. The chemical structure is a two-dimensional (2D) object, which can be represented as a graph consisting of vertexes (atoms) and edges (bonds). Thus, a straightforward method for comparing two compounds is graph comparison, or detecting common (isomorphic) subgraphs in two graphs.

In practice, however, the comparison of bit-represented vectors, which is not a graph comparison, has been utilized as a common method for searching similar compounds in a chemical database.[15] In this method the information about a compound structure is reduced to a concatenation of several hundreds of bits.[16] A numerical vector method[17,18] and a fingerprint method[19] have also been used as a mathematical extension of the bit-comparison method. In contrast, comparing two compounds directly as graph objects by using graph theoretical methods is one of the major categories of applications that need further developments and refinements. Especially, it is critical to define an appropriate measure of compound similarity for any graph comparison method to be biochemically meaningful.[20,21] The representation of compounds as graphs seems more accurate and more effective to capture important aspects of compound similarities[22] rather than other representations of compounds such as SMILES.[23,24] Recently, advances have been made in the graph similarity search algorithms by taking mathematical or chemical heuristics into account.[25−27] These algorithms may be of practical use in the field of chemical software systems. On the other hand, graph comparison methods have a fundamental difficulty; the graph isomorphism problem is NP-hard, and the computational time involved will increase exponentially for larger biochemical compounds.

In this study we have developed a suite of new computational tools, named SIMCOMP, to annotate an atomic environmental property for each atom of a biochemical compound, to rapidly identify common substructures between two compounds on the basis of a graph comparison method, and to evaluate statistical significance of similar substructures. Biochemical dialects of compounds are sometimes useful to identify common properties of compounds,[28−30] and we first try to include biochemical information into the representation of atoms, by distinguishing the same atoms under different environments. This effectively increases the number of vertex types and reduces the limitation of 2D graph utilization. In addition, we introduce several heuristics into the algorithm of similarity calculations. Thus, we could decrease the exponential difficulties of graph comparison methods to the practical level that can be tolerated, while holding high accuracies for graph similarities found.

Our method is applied to comparison and classification of about 10 000 compounds, mostly metabolic compounds, in KEGG. In particular, we perform a pathway-oriented clustering, which reveals highly conserved modules of metabolic pathways, consisting of successive reaction steps involving similar chemical compounds. Because the relationships between genomic contexts (e.g., operon structures) and pathway modules are already well identified by a number of studies and collected in KEGG as ortholog group tables,[31] it is natural for us to examine any correspondence between chemical information and genomic information, how well pathway modules identified by genomic contexts correspond to those identified by chemical contexts. This is a new type of network analysis, integrating both chemical and genomic information for understanding molecular interaction networks.

Materials and Methods

Chemical Compound Data. We have used chemical compound data in the COMPOUND section of the KEGG/LIGAND database (version 20.0 + update 2002/03/26),[32,33] which is maintained in the ISIS/Oracle database system. The total number of compounds with chemical structures is 9383, roughly classified, according to the source, into 977 drug-related compounds, 2649 phytochemical compounds (secondary metabolites in plants), and 5757 metabolites and other compounds originating mostly from the KEGG metabolic pathways and/or the enzyme nomenclature (EC number classification). We consider each chemical structure as a labeled graph with atoms (or atom types) as its vertexes and covalent bonds as its edges, excluding hydrogen atoms. We do not consider any 3D features and do not discriminate chirality. Some KEGG compounds are described in a generic form or a polymeric

(12) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536−540.
(13) Conte, L. L.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **2002**, *30*, 264−267.
(14) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH: A hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093−1108.
(15) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(16) Allen, F. H.; Kennard, O. 3D search and research using the Cambridge structural database. *Chem. Des. Autom. News* **1993**, *8*, 1 and 31−37.
(17) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(18) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(19) James, C. A.; Weininger, D.; Delany, J. Daylight Theory Manual 4.71; Daylight Chemical Information Systems, Inc.: Irvine, CA, 2000.
(20) Willett, P. Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. *J. Mol. Recognit.* **1995**, *8*, 290−303.
(21) Miller, M. A. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discovery* **2002**, *220*, 220−227.
(22) Arita, M. Graph modeling of metabolism. *J. Jpn. Soc. A. I.* **2000**, *15*, 703−710.
(23) Weininger, D. SMILES 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(24) 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.
(25) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.
(26) Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of graph similarity using maximum common edge subgraphs. *Comput. J.* **2002**, *45*, 631−644.
(27) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305−316.
(28) Mavrovouniotis, M. L. Group contributions for estimation standard Gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.* **1990**, *36*, 1070−1082.
(29) Mavrovouniotis, M. L. Estimation of standard Gibbs energy changes of biotransformations. *J. Biol. Chem.* **1991**, *266*, 14440−14445.
(30) Forsythe, R. G., Jr.; Karp, P. D.; Mavrovouniotis, M. L. Estimation of equilibrium constants using automated group contribution methods. *Comput. Appl. Biosci*. **1997**, *13*, 537−543.
(31) Fujibuchi, W.; Ogata, H.; Matsuda, H.; Kanehisa, M. Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. *Nucleic Acids Res.* **2000**, *28*, 4029−4036.
(32) Goto, S.; Nishioka, T.; Kanehisa, M. LIGAND: chemical database for enzyme reactions. *Bioinformatics* **1998**, *14*, 591−599.
(33) Goto, S.; Okuno, Y.; Hattori, M.; Nishioka, T.; Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **2002**, *30*, 402−404.

form, such as primary alcohol (R−OH) or starch ($\{C_{12}H_{20}O_{11}\}_n$), which is often necessary to better represent metabolic pathways. We treat these compounds by the following rules: (1) the R group is just taken as "R" atom, that is, as if R were the 69th atom type (in addition to the 68 types described in the Results), and (2) the degree of polymerization n is taken as 1, which means any polymeric structures degenerate to corresponding monomers.

Definition of Graph Features. The problem of finding chemical compound similarities is a graph comparison problem. Our approach to finding common (isomorphic) subgraphs is essentially the same as the traditional association graph method,[34,35] which provides one of the efficient solutions for the graph isomorphism problem. Here we summarize the terminology for relevant graph features.

(1) Maximum Clique (MCL). A vertex-labeled graph consists of the set of vertexes V and the set of edges E, and is denoted by $G(V,E)$. A clique of graph G is defined as a complete subgraph in G. The maximum clique in graph G is the clique of G whose cardinality is not smaller than that of any other clique in G. The maximum clique of graph G is denoted as MCL(G).

(2) Maximal Common Subgraph (MCS) and Simply Connected Common Subgraph (SCCS). A subgraph of graph G is a new graph obtained from G by deleting some edges and vertexes. A common subgraph of G_1 and G_2, CS(G_1,G_2), is a graph which is isomorphic to a subgraph of both G_1 and G_2. The maximal common subgraph of G_1 and G_2, MCS(G_1,G_2), is the CS(G_1,G_2) whose cardinality is not smaller than that of any other CS(G_1,G_2). A simply connected common subgraph, SCCS(G_1,G_2), is a CS(G_1,G_2) within which each vertex is connected to at least one other vertex. The MCS(G_1,G_2) must be a series of SCCS(G_1,G_2)'s.

(3) Association Graph (AG). The graph product GP(V,E) of two graphs $G_1(V_1,E_1)$ and $G_2(V_2,E_2)$ is a new graph defined on the vertex set $V = V_1 \otimes V_2$ (a Cartesian product of V_1 and V_2) and the set of edges $E = V \otimes V$. The association graph AG(V,E) defined here is one of the graph products with the following adjacency conditions. Any $e(v_{ij},v_{st}) \in E$ is considered to be adjacent (1) if $v_{1i} \in V_1$ is adjacent to $v_{1j} \in V_1$ in the original graph G_1 and $v_{2s} \in V_2$ is adjacent to $v_{2t} \in V_2$ in the original graphs G_2, or (2) if v_{1i} is not adjacent to v_{1j} and v_{2s} is not adjacent to v_{2t}.

Clique Finding in the Association Graph. The association graph AG made by the previous definition possesses all possibilities of vertex matches between two initial graphs G_1 and G_2; namely, a clique in AG corresponds to a common subgraph between G_1 and G_2. Thus, the largest clique based on the number of matching vertexes becomes the largest match of our interest. Consequently, the initial problem of finding the MCS(G_1,G_2) can be reduced to the problem of finding the MCL(AG). We use this association graph method only to obtain an initial candidate set of maximally matching atoms (see Results).

Results

Atom Types with Different Environments. The structure of a chemical compound is a collection of atoms (vertexes) that are connected by covalent bonds (edges). In this study, any 3D structural information of edges is not implemented; that is, we use the graph representation containing only the 2D information about vertexes and vertex connectivities in chemical compounds. Although we discard 3D atomic coordinates, which of course are not available for most compounds, we take into account physicochemical environmental properties of atoms by assigning well-detailed vertex labels. The same atoms in chemical compounds may thus be distinguished by different labels, because they represent different physicochemical properties in accordance with their spatial and chemical situations. For instance, a carboxyl carbon (R−(C═O)−OH) and an aldehyde carbon (R−(C═O)−H) are very similar and have the same atomic bond skeleton (X−(C═X)−X), which is one of the most basic building blocks of larger molecules. However, these two types of carbons are obviously different from the viewpoint of organic reactions because of the difference in reactivities. It is a well-known fact that an aldehyde carbon is more active on a nucleophilic addition reaction, while a carboxyl carbon usually has a nucleophilic substitution reaction activity. Therefore, it is reasonable that we discriminate these two types of carbon when comparing molecules.

Such atom-typing has commonly been utilized in chemoinformatics. Here, we also introduce the vertex labeling function $p(v)$ into the graph representation of chemical compounds. The labeling function should reflect the environmental features of atoms and is based on the examination of the following: (1) whether the atom is included in a ring structure, (2) what types of bonds are connected to the atom, for example, single, double, triple and aromatic bonds, and (3) what atoms are adjacent and, if needed, what atoms are further adjacent to the adjacent atoms.

This labeling system is very simple and can be generated computationally on the basis of the connection patterns of atoms and the functional groups that they belong to and without any other supervisor knowledge. Hence, each atom of all chemical compounds in KEGG could be assigned new labels automatically from their initial graphs stored in the MDL/MOL file format. Figure 1 shows the list of new labels and corresponding atomic environments as well as the numbers of instances found in the KEGG compounds.

Thus we distinguish carbon into 23 types, nitrogen into 16 types, oxygen into 18 types, sulfur into 7 types, and phosphorus into 2 types. The total number of new atom types is 68 including two more types for halogens and the rest. In this new labeled graph representation of chemical compounds, carboxyl carbon (R−(C═O)−OH) and aldehyde carbon (R−(C═O)−H) are now considered different, C6a and C4a, respectively. This representation is thus able to distinguish functional groups and should be able to identify similarities and differences of biochemical features of chemical compounds. For example, as illustrated in Figure 2, although 3-hydroxypropanoate and 3-oxopropanoate are very similar and have the same graph topology, the difference between these two compounds can be detected by referring to differently labeled vertexes indicating that 3-oxopropanoate has an aldehyde group.

Weighting of Atom Type Matches. The problem of finding the maximal common subgraph (MCS) in two graphs is known to be solved by finding the maximal clique (MCL) in the so-called association graph consisting of the products of vertexes from two graphs as its vertexes. In a conventional method, each vertex of the association graph is weighted as only one or zero, called all-or-none weighting here, depending on whether two vertexes from the original graphs do or do not match. However, this type of weighting scheme is too strict for our representation where 68 atom types obviously share one of the seven categories of atomic species. A simple weighting scheme adopted here, called loose weighting, allows partial matches for the same atom species with different environments, such as carboxyl carbon and aldehyde carbon.

(34) Kuhl, F. S.; Crippen, G. M.; Friesen, D. K. A combinatorial algorithm for calculating ligand binding. *J. Comput. Chem.* **1984**, *5*, 24−34.
(35) Takahashi, Y.; Maeda, S.; Sasaki, S. Automated recognition of common geometrical patterns among a variety of three-dimensional molecular structures. *Anal. Chim. Acta* **1987**, *200*, 363−377.

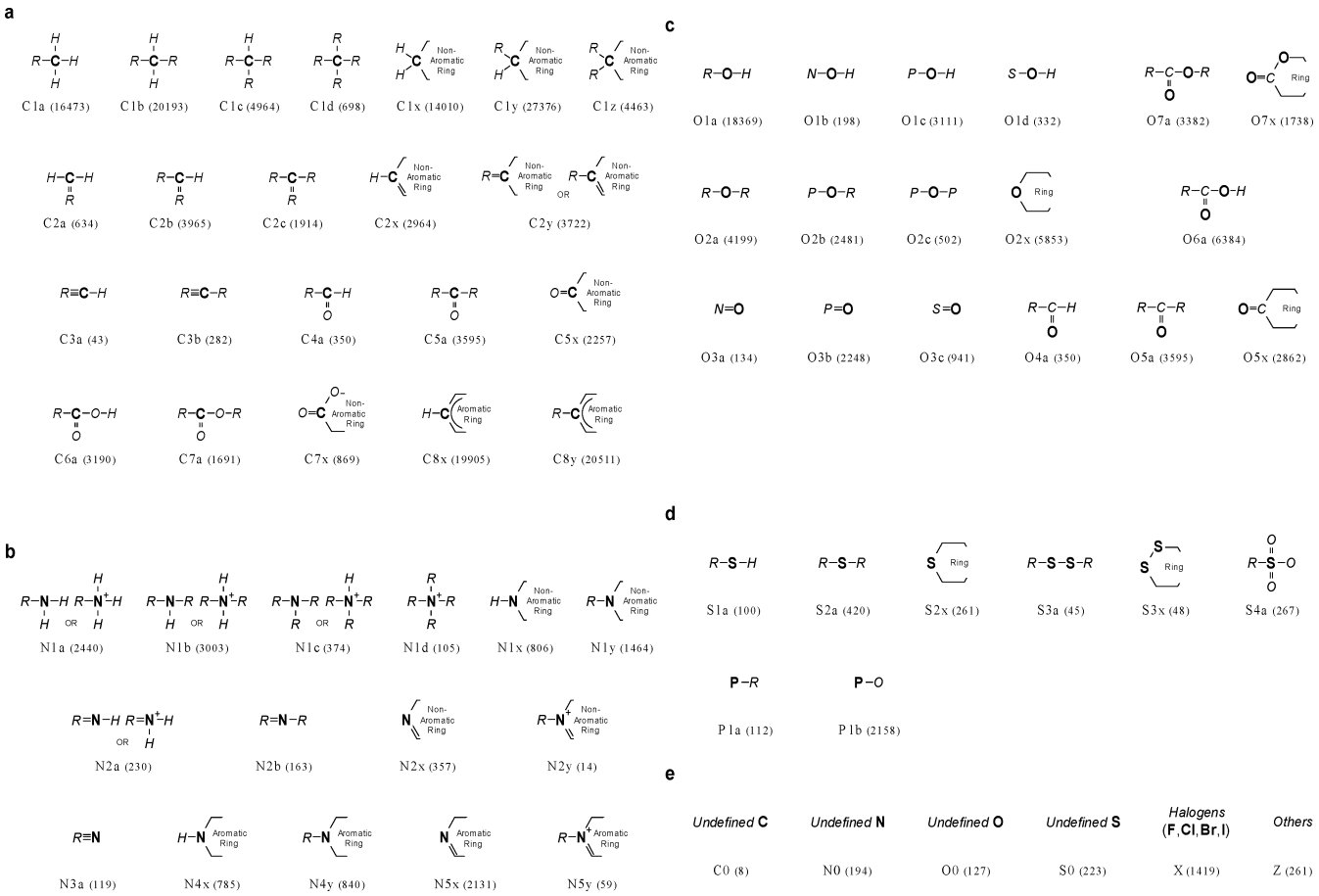

Figure 1. List of 68 atom types that distinguish environmental classes. The atom-type codes are shown for carbon (C) in diagram a, nitrogen (N) in b, oxygen (O) in c, both sulfur (S) and phosphorus (P) in d, and the rest in e. In each diagram H is a hydrogen atom and R is an atomic group larger than a simple hydrogen atom including a ring. In some cases, such as O6a, O7a, O7x, S3a, or S3x, atom-type codes are assigned to plural target atoms. The last category e is miscellaneous containing any C, N, O, or S with no suitable class in a, b, c, or d. A halogen is labeled as X, and other atoms are reduced into Z. The observed frequencies of each atom type in our dataset are also shown in parentheses.

The scheme is formulated as follows. Given two graphs $G_1(V_1,E_1)$ and $G_2(V_2,E_2)$, the vertex v_{ij} of the association graph AG(V,E) is induced from two vertexes $v_{1i} \in V_1$ and $v_{2j} \in V_2$ and is weighted as:

$$w(v_{ij}) = \begin{cases} 1, & \text{if } p(v_{1i}) = p(v_{2j}), \\ c, & \text{if } p(v_{1i}) \neq p(v_{2j}) \text{ and } a(v_{1i}) = a(v_{2j}), \\ 0, & \text{otherwise} \end{cases}$$

Here, the function $a(v)$ returns the atom species of vertex v, and c is the constant value between 0 and 1. Of these three statements, the first and the third ones are counterparts of the all-or-none rule in the traditional association graph method. Here we have introduced the second statement, which allows the pairing of different atom types when the atom species is the same. Through this newly weighted association graph AG, we can still define the maximal common subgraph MCS(G_1,G_2) as the maximal clique MCL(AG). In the current implementation of our SIMCOMP program, we first obtain all cliques with the maximum number of vertexes by the clique finding algorithm shown below, and then calculate the sum of weights

$$\sum_{v \in \mathrm{MCL(AG)}} w(v)$$

for each clique to select the largest weighted one.

The parameter c is an adjustable parameter. As c goes to 0, the computational result approximates to that of the conventional all-or-none weighting rule. When c turns to 1, it will become the same as the result without the complicated vertex labels. In this study, we chose $c = 0.5$ as an intermediate degree of atom matches. The distinction between the all-or-none type and the loose type of weighting is illustrated in Figure 2. The SIMCOMP program with the all-or-none weighting detects only the common structure (1), but with the loose weighting of $c = 0.5$ it detects the common structure (2) as well.

Improvements of the Clique Finding Algorithm. The clique finding of a given graph is a well-studied problem and it is known to be combinatorially explosive in nature. Our implementation of the clique finding is a modified version of the Bron−Kerbosch algorithm.[36] Since the association graph AG(V,E) is generated only for the matching vertexes in the initial gaphs, the number of vertexes in AG is much larger under the loosely weighted condition than the all-or-none condition, and the calculation based on this algorithm does not finish within a practical time for many compound pairs in our database. Thus, we need to incorporate better heuristics into the calculation.

First, we simply stop the calculation of clique finding after a reasonable number of recursion steps in a recursive implementa-

(36) Bron, C.; Kerbosch, J. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM* **1973**, *16*, 575−577.

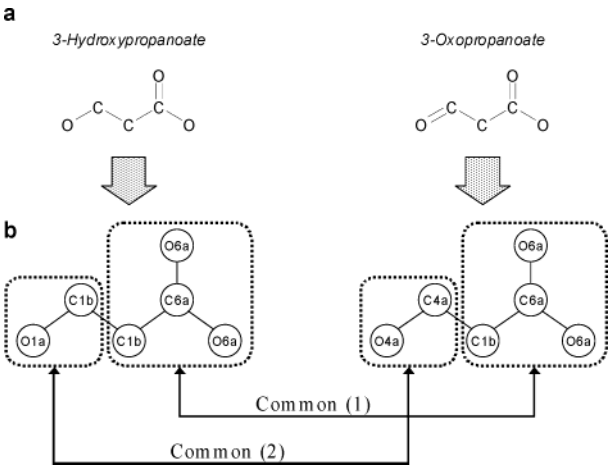

Figure 2. Conversion of atoms into atom types distinguishing environments. (a) Initial graphs of chemical compounds obtained from KEGG/LIGAND; in this case, 3-hydroxypropanoate and 3-oxopropanoate. (b) Conversion into a more complicated graph whose vertexes are labeled by the proper atom types listed in Figure 1. The common subgraphs that should be detected by our method are also shown here. Under the all-or-none weighting condition the common substructure (1) is detected with the normalized similarity score of $4/(6 + 6 - 4) = 0.5$. Under the loose weighting condition the common substructure (2) can also be identified, and the whole structures of two compounds are found to have the same topology with the normalized similarity score of 1.

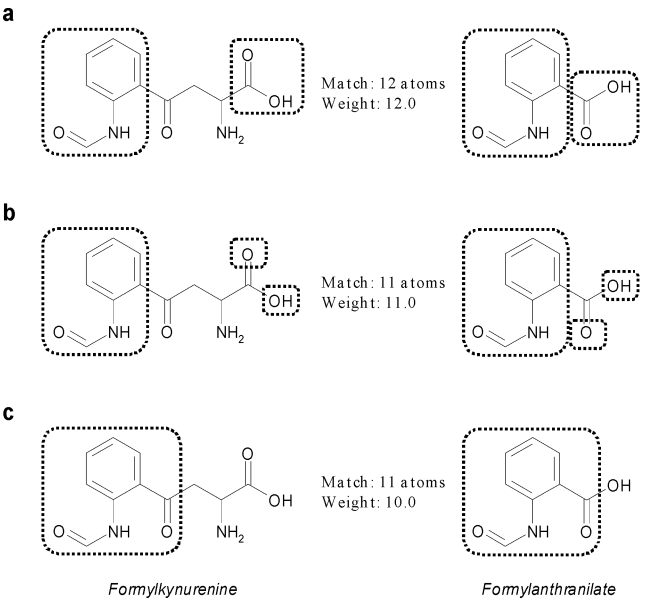

Figure 3. Heuristics of maximizing simply connected common subgraphs. (a) Rigorous clique-finding procedure detects this best solution, that is, the maximal common subgraph (MCS) between formylkynurenine and formylanthranilate. (b) Suboptimal solution, that is, a quasi-MCS after calculating up to the given number of steps ($R_{max} = 15\,000$). (c) The result of eliminating small SCCSs ($S_{min} = 2$) and maximizing the other larger SCCSs. This solution is mathematically less optimal than either a or b, but biochemically meaningful.

tion of the Bron−Kerbosch algorithm and obtain a candidate set of MCLs (maximal cliques), that is, MCSs (maximal common subgraphs) as well. Then we start to search better common subgraphs, called quasi-MCSs, from the candidate set. In this second optimization step we eliminate small SCCSs (simply connected common subgraphs) whose cardinality is smaller than a given threshold, and extend only other larger SCCSs. The SCCSs with small cardinality are frequently found as noises around the conserved structure of two compounds, such as separate matches of single atoms. Mathematically those separate matches should be considered to obtain the MCS, but the quasi-MCS without considering them may be biochemically meaningful. After the elimination of those small SCCSs, we extend the other SCCSs one by one greedily until no more atom pairs can be included. Finally we obtain the quasi-MCS(G_1,G_2).

The procedure outlined above thus contains heuristics summarized below: (1) to suspend the clique finding procedure at the number of recursion steps R_{max}, at most, (2) to eliminate any small SCCSs whose cardinality is lower than S_{min}, and (3) to extend the other SCCSs greedily while any candidate exists.

Our two-step optimization procedure is controlled by the two cutoff parameters, R_{max} for termination of the usual clique finding algorithm and S_{min} for consideration of the greedy search around each of the SCCSs found. In this paper we chose $R_{max} = 15\,000$ and $S_{min} = 2$, after several preliminary experiments on computing chemical compound similarities in the KEGG database.

The heuristics introduced here not only made the computation more efficient but also made it possible to capture biochemically meaningful features, as illustrated in Figure 3. When formylkynurenine and formylanthranilate are compared by the rigorous clique finding algorithm, the maximal common substructure is identified as structure **a**. However, in our heuristic procedure of discarding small SCCSs with size one from the solution after 15 000 steps (structure **b**) and searching for larger SCCSs, the final result was structure **c**, which is less optimal than **a** or **b** but is more appropriate from the biochemical standpoint. This is because there exists an enzymatic reaction between these two compounds (EC: 3.7.1.3) where formylkynurenine is divided into formylanthranilate and L-alanine, and the common substructure **c** does represent this reaction. In many other cases that we examined, we obtained relatively reasonable solutions with $S_{min} = 2$ especially for closely related compound pairs.

Normalized Score for Compound Similarity. The maximal common subgraph MCS(G_1,G_2) is obtained by maximizing the number of matched atom types, which is a raw score that depends on the sizes of the original graphs G_1 and G_2. We introduce a normalized score, utilizing one of the most popular measures, the Jaccard coefficient,[37,38] also known as the Tanimoto coefficient.[39,40] It is the ratio of the size of the common substructure (AND graph) divided by the size of the nonredundant set of all substructures (OR graph). The OR graph consists of one isomorphic copy of the subgraph existing in both graphs and all other subgraphs existing in either graph, and is defined as $G_1 + G_2 - \mathrm{MCS}(G_1,G_2)$.

Thus, the Jaccard coefficient JC(G_1,G_2) that is the cardinality of the common subgraph divided by the cardinality of the nonredundant subgraph can be written as:

$$\mathrm{JC}(G_1,G_2) \equiv \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|} = \frac{|\mathrm{MCS}(G_1,G_2)|}{|G_1 + G_2 - \mathrm{MCS}(G_1,G_2)|} = \frac{|\mathrm{MCS}(G_1,G_2)|}{|G_1| + |G_2| - |\mathrm{MCS}(G_1,G_2)|}$$

(37) Jaccard, P. The distribution of the flora of the alpine zone. *New Phytol.* **1912**, *11*, 37−50.

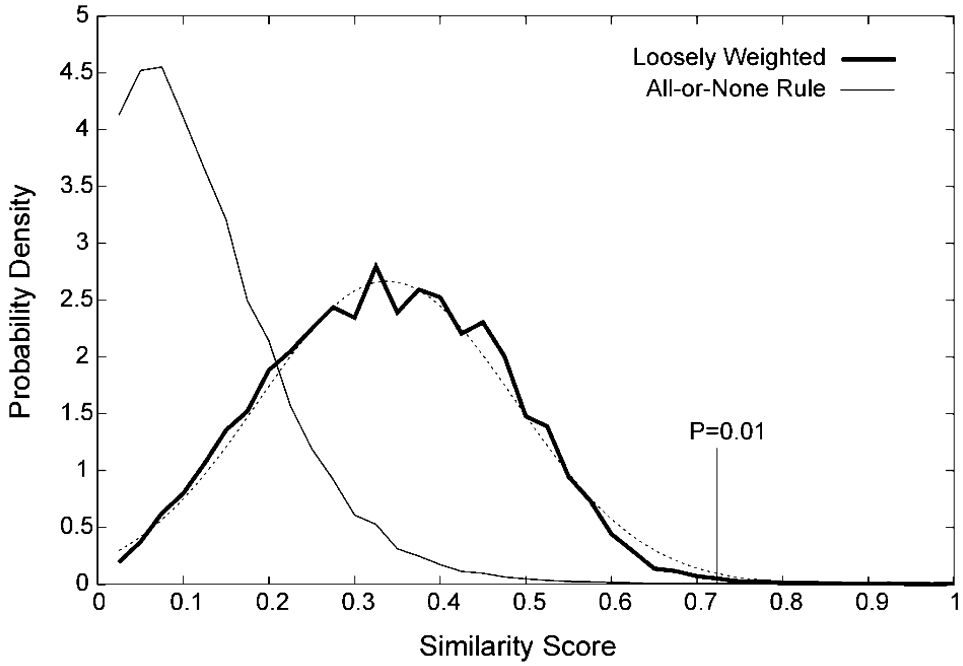

Figure 4. Distribution of normalized similarity scores for all possible pairs of chemical compounds in KEGG. The thick line is the probability density distribution with the loose weighting condition, and the thin line is that for the all-or-none weighting condition. Here the thick line can be fitted with a normal distribution, drawn in a dashed line, whose statistical parameters are $\mu = 0.338$ and $\sigma = 0.150$. According to this normal distribution P-value = 0.01 for the right tail corresponds to score = 0.723, as indicated in the figure.

where the notation $|X|$ is used for the cardinality of graph X. Because we search for quasi-MCS(G_1,G_2), the Jaccard coefficient is approximated by:

$$\mathrm{JC}(G_1,G_2) \approx \frac{|\mathrm{qMCS}(G_1,G_2)|}{|G_1| + |G_2| - |\mathrm{qMCS}(G_1,G_2)|}$$

The normalized similarity score JC ranges from 0 to 1, where 0 represents the absence of any common substructure and 1 means that two compounds are identical.

Comparison of All Compound Pairs in KEGG. We calculated the normalized similarity scores for all possible pairs of chemical compound structures in the KEGG dataset using the SIMCOMP program under the loose weighting condition. The distribution of 44,015,653 similarity scores among 9383 compounds is shown in Figure 4. The statistical distribution that best approximates this distribution is found to be a normal distribution, also drawn in Figure 4. This probability density function is formulated as:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Here, μ is the average of all similarity scores, and σ is the standard deviation of this distribution. The probability $P(s > S)$ of observing by chance the score s that is greater than S is given by:

$$P(s > S) = \int_S^\infty F(x)\mathrm{d}x = \frac{1}{\sigma\sqrt{2\pi}} \int_S^\infty \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x$$

and it is referred as the P-value. From this equation, we estimated the threshold of the similarity score in order to best discriminate biochemically meaningful compound pairs. For our particular dataset, we chose P-value = 0.01, or the level of confidence of 99%; thus, the proper threshold is $S = 0.723$.

Figure 4 also shows the distribution of similarity scores for all possible KEGG compound pairs with the all-or-none weighting, which requires perfect matching of 68 atom types. Few common substructures were found for most compound pairs with the all-or-none weighting as indicated by a skewed distribution similar to the binomial distribution, in contrast to the normal distribution in the loose weighting.

Clustering of All Compounds in KEGG. After calculating similarity scores of all possible compound pairs in our dataset, we performed the complete-linkage cluster analysis with the threshold similarity score of 0.723 (the degree of confidence 99%). Consequently, the total number of clusters found was 3970, consisting of 1871 singletons and 2099 non-singletons, and the maximum size cluster contained 64 compounds. As shown in Figure 5, the size distribution exhibits the "small world" nature[41] approximately following the power-law distribution.

By examining constituent members of each cluster in more detail, we found that clusters with large numbers of metabolites were often associated with specific compound families. The top 10 largest clusters are listed in Table 1, and for each of them a representative structure is shown in Figure 6 together with the common substructure. Obviously, many of the largest clusters

(38) Watson, G. A. An algorithm for the single facility location problem using the Jaccard metric. *SIAM J. Sci. Stat. Comput.* **1983**, *4*, 748−756.
(39) Willett, P.; Winterman, V.; Bawden, D. Implementation of nearest-neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.
(40) Willett, P.; Barnard, J.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(41) Barabasi, A. L.; Albert, R. Emergence of scaling in random networks. *Science* **1999**, *286*, 509−512.

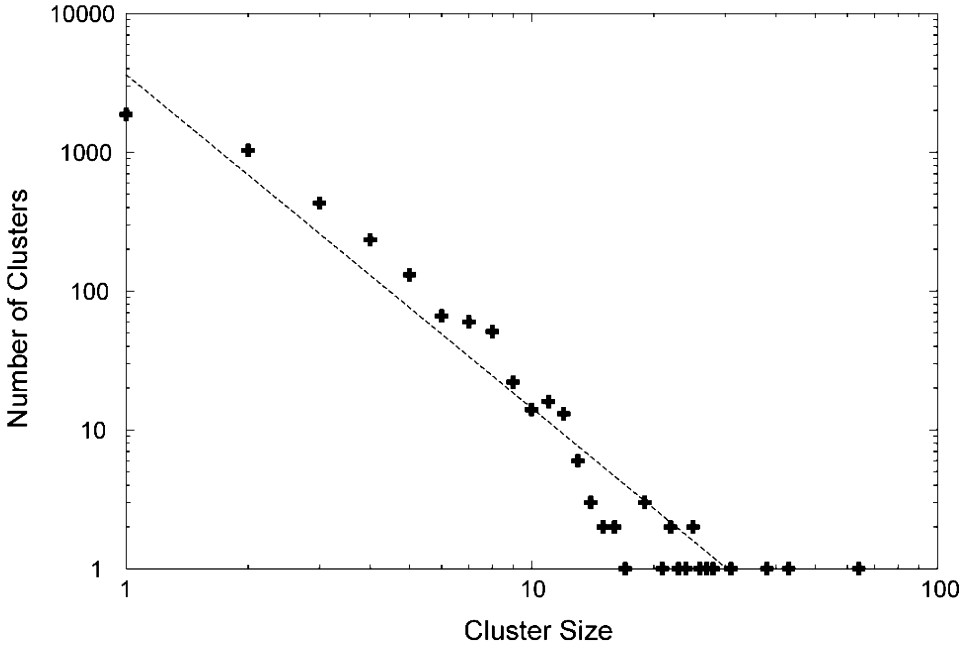

Figure 5. Size distribution of similar compound clusters that are identified by the complete linkage analysis with the threshold similarity score of 0.723 (the degree of confidence is 99%). In this log−log plot, the horizontal axis is the cluster size or the number of compounds belonging to the cluster, and the vertical axis is the number of clusters with a given size. The dashed line is the regression line, indicating that the size distribution of clusters approximately follows the power-law, $P(k) \propto k^{-\gamma}$, with $\gamma = 2.41$.

Table 1. Top Ten Largest Clusters of Similar Chemical Compounds

				KEGG pathways map numbers[a]						
no.	size	common formula	description of members	C	L	N	AA	CC	second	AtR
1	64	C_6O_6	hexose, its uronic acid, glycoside	10, 30, 52				**500**		
2	43	C_6O_5	ketohexose, aldohexose, aldarate	30, **40**, 51, 52, 53						
3	38	C_5O_5P	ribose and phosphoric acid group of nucleic acids							**970**
4	31	C_6O_8P	phosphorylated hexose	**51**, 52		520				
5	28	C_5O_5	ketopentose, hexose lactone	40, **53**						
6	27	C_9O	containing a cinnamate skeleton				350, **360**		940	
7	26	C_5O_4	aldopentose, pentoside	**40**		520				
8	25	C_{10}	containing a menthol skeleton						900	
9	25	$C_{27}O$	containing a cholesterol skeleton		**100**					
10	24	C_8O_6N	*N*-acetylated hexosamine					**530**		

[a] The pathway map numbers are simplified; for example, 40 stands for map00040 in KEGG. The most frequently observed pathways are shown in bold. Abbreviations for the pathway categories are: C, carbohydrate metabolism; L, lipid metabolism; N, nucleotide metabolism; AA, amino acid metabolism; CC, metabolism of complex carbohydrates; second, biosynthesis of secondary metabolites; and AtR, aminoacyl-tRNA synthesis.

consist of sugar-related compounds; especially the clusters 1, 2, and 10 have common skeletons of hexoses. As a matter of fact, the clusters 1 and 2 become connected into a single cluster at the similarity threshold = 0.6, and 1 and 2 and 10 are grouped into one cluster at the threshold = 0.5. The cluster 4 is also a group of hexose-related compounds, but it is separated from others until the threshold is less than 0.4. The clusters 5 and 7 are related to pentoses, but they are distinct groups even the threshold score is lowered to 0.4. These characteristics may arise from the nature of the complete linkage analysis, that is any pair within the cluster must have a similarity score above the given threshold. In any event, we could identify chemically distinct groups at the high-confident threshold = 0.723, which are likely to represent biochemically meaningful groups as summarized in Table 1.

We have also noticed that most of the top 10 largest clusters are highly correlated with specific metabolic pathways (Table 1). Here, the correspondence between a cluster and a pathway is defined by the number of compounds within a cluster that can be assigned to a specific pathway map in KEGG. For instance, all compounds included in the cluster 9 are associated with the metabolic pathway of sterol biosynthesis, whose accession number in KEGG is map00100. The cluster 6 is strongly connected with phenylalanine (map00360) or tyrosine (map00350) metabolism. Most of the other top ranking clusters are correlated with carbohydrates that appear ubiquitously in many metabolic pathways in KEGG, especially, map00040 (pentose and glucuronate interconversion), map00052 (galactose metabolism), and map00053 (ascorbate and aldarate metabolism). The total number of compounds that can be mapped to KEGG metabolic pathways was 2294, roughly a quarter of 9383 compounds in our dataset.

Clustering of Compounds within KEGG Pathway Maps. The cluster analysis of all 9383 compounds revealed the global tendency of similar compounds appearing in the same KEGG metabolic pathway maps. An obvious next question is whether those similar compounds are also related to specific reaction steps when each pathway map is examined in more detail. We thus checked the connectivity of compounds along the reaction steps by mapping similar compound clusters onto KEGG

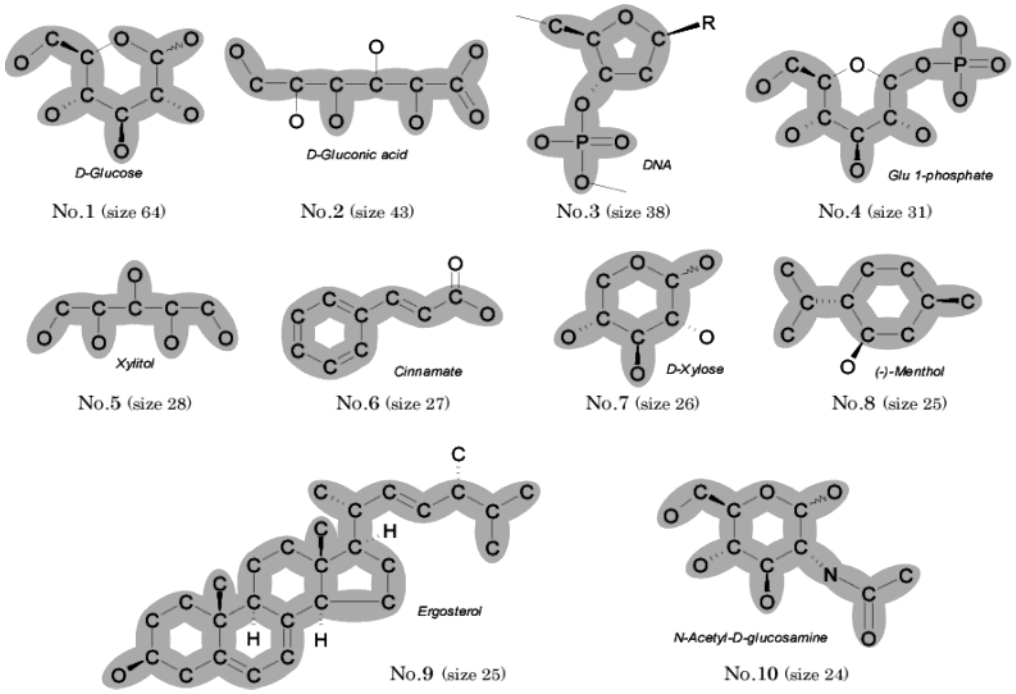

Figure 6. Common structures of the top 10 largest clusters. For each cluster a most representative compound is shown with its name, and the common structure is indicated in gray. The cluster size is shown in parentheses. The clusters 1, 2, 4, and 10 belong to the group of hexoses and derivatives, and their common structures are very similar. In fact, the four representative chemical compounds in this figure have high similarity scores each other. Apparently, the complete linkage method makes these clusters separated at the threshold score of 0.723, as well as the pentose-related clusters 5 and 7.

pathway maps. Although the above result of clustering all 9383 compounds could be used for this purpose, we also performed the cluster analysis of 2294 compounds that were already known to appear in the KEGG pathway maps. This pathway-oriented clustering was carried out in the same way as above, the complete-linkage clustering with the threshold score of 0.723.

The result of mapping compound clusters onto each KEGG pathway is summarized in Table 2 for both types of cluster analyses. There was a definite tendency that similar compound clusters corresponded to localized regions of the pathway maps, indicating that compounds of high structural similarities are also likely to be linked with high connectivities on the reaction steps. With the pathway-oriented clustering, most of the KEGG metabolic pathway maps could be divided into several parts of chemical compound clusters more plainly than the case of all compounds. In addition, some of the metabolic pathways had larger components of pathway clusters, and the correspondences between compound clusters and pathway maps became clearer.

As an example, the result of analyzing the KEGG metabolic pathway map for pentose and glucuronate interconversions (map00040) is shown in Figure 7. Four compound clusters were identified by the pathway-oriented clustering as indicated in Figure 7a and the consensus structure of each compound cluster is shown in Figure 7b. It is obvious that this map is largely separated into two parts; one is the pentose-related region (clusters B and D) and the other is the glucuronate-related region (cluster A). Cluster C is located between B and D, for any member of C is a phosphorylated product of B or D as shown in Figure 7b. Here the consensus structure is the common skeleton of atoms identified by the atom alignment in SIMCOMP, namely, without considering atomic environmental properties.

Correlation of Compound Clusters and Operon Structures. One of the main objectives of this study is to find, if any, empirical relationships between chemical information and genomic information in the metabolic pathways. The chemical information is derived from the cluster analysis of chemical compounds and the pathway-oriented clustering as described above. The genomic information considered here is taken from the KEGG ortholog group tables,[31,42] which contain the information about orthologous sets of enzyme genes that constitute specific pathways and also about enzyme gene clusters (possible operons) in selected genomes. The correlation is assessed by projecting both chemical compound clusters and enzyme gene clusters onto each KEGG metabolic pathway map and enumerating the number of compounds in the intersection of these two types of clusters. Thus, the chemical compounds in the intersection would exhibit three significant features: high structural similarity, connectivity or reactivity of compounds along the pathways, and genomic association of enzymes catalyzing reactions between those compounds.

The last two columns of Table 2 show the number and the maximum size of intersection clusters that we obtained. The enzyme gene clusters (operon structures) were correlated well with the pathway-oriented compound clusters in almost all KEGG pathway maps, but the intersection was usually small. The largest intersection was found in map00040 for pentose and glucuronate interconversions, which is illustrated in Figure 8. The region A is the cluster of similar compounds (glucuronates) shown in Figure 7a. The region E is the cluster of enzyme genes, which actually contain three operon-like structures in certain genomes. The first operon-like structure (such as in *Yersinia pestis*[43]) consisting of EC 4.2.1.7, EC 1.1.1.58, and EC 5.3.1.12 and the second operon-like structure (such as in *Brucella melitensis*[44]) consisting of EC 4.2.1.8, EC 1.1.1.57,

(42) 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.

and EC 5.3.1.12 are found within the compound cluster A, and the third operon-like structure (such as in *Bacillus subtilis*[45]) consisting of EC 5.3.1.17, EC 1.1.1.125, EC 2.7.1.45, EC 4.1.2.14, and EC 4.1.3.16 partially overlaps with the compound cluster A. Thus, the shaded area in Figure 8 represents a highly conserved pathway module, which represents both chemical similarity of compounds and genomic association of enzymes. There were also similar but smaller intersections in map00040 where an enzyme gene cluster (such as in *Escherichia coli*[46] and *Salmonella*[47]) was found to overlap compound clusters B (pentoses) and C (phosphorylated products). All such relationships between gene clusters and compound clusters in map00040 are listed in Table 3.

Discussion

Integration of Chemical and Genomic Information. The correlation between the genomic association and the pathway connectivity is already well-known; a set of enzyme genes encoded in an operon often corresponds to a set of enzymes catalyzing successive reaction steps in a specific metabolic

Table 2. Numbers of Compound Clusters and Enzyme Gene Clusters Found

path-way		total		all compounds			by pathway			by EC	
		CPD	EC	Num1	Num2	Max	Num1	Num2	Max	NumE	MaxC
C	map00010	32	12	20	10	3	16	9	5	6	3
	map00020	22	11	15	5	4	13	5	4	3	4
	map00030	30	13	15	7	5	12	6	7	5	4
	map00040	50	21	20	10	8	19	4	16	4	9
	map00051	50	17	22	15	5	24	9	7	4	3
	map00052	41	14	26	7	7	22	8	11	5	6
	map00053	31	4	11	7	10	13	9	5	1	4
	map00620	28	4	17	5	5	17	3	8	0	-
	map00630	43	6	24	11	4	25	8	7	1	3
	map00640	36	9	26	9	3	25	7	5	2	2
	map00650	40	5	23	9	5	19	9	7	1	5
E	map00190	12	7	10	2	2	10	2	2	2	2
	map00680	26	3	21	4	2	20	5	2	2	2
	map00910	25	5	17	6	3	20	4	2	1	2
	map00920	59	6	44	6	3	49	3	2	1	2
L	map00061	36	8	10	4	7	14	6	5	6	5
	map00062	30	7	16	9	3	12	8	5	7	5
	map00071	51	7	26	13	3	21	9	8	5	8
	map00100	66	6	31	13	7	50	9	3	1	3
N	map00230	88	21	45	18	8	56	18	9	3	5
	map00240	59	24	31	13	6	37	12	4	8	4
	map00520	33	7	12	7	10	14	5	8	3	5
AA	map00251	28	10	23	4	3	22	5	3	1	2
	map00252	27	6	20	6	3	21	4	4	1	3
	map00260	53	17	36	13	4	34	14	5	6	3
	map00271	20	2	14	4	3	13	4	4	1	2
	map00272	23	2	14	4	4	17	4	3	0	-
	map00280	36	7	20	7	4	19	7	6	5	4
	map00290	23	10	13	6	5	17	3	4	3	4
	map00300	33	12	20	9	3	16	8	5	4	2
	map00330	70	7	47	14	5	50	8	11	1	3
	map00340	45	13	26	10	4	27	11	4	5	3
	map00350	82	6	35	19	11	37	18	6	2	3
	map00360	31	9	17	7	9	19	6	4	1	2
	map00400	26	16	20	5	3	17	3	5	3	4
	map00220	33	16	28	5	2	28	4	3	1	2
oAA	map00410	30	5	23	5	4	21	5	4	3	2
CC	map00500	53	12	30	9	8	24	12	7	5	4
	map00530	31	3	17	7	5	11	7	7	1	3
	map00540	16	8	15	0	1	15	0	1	0	-
	map00550	37	7	27	6	4	29	5	3	1	2
CL	map00561	70	5	43	16	5	49	13	4	1	2
CoV	map00730	15	4	11	2	4	9	3	4	1	2
	map00740	19	6	15	3	3	13	5	3	3	3
	map00760	23	2	15	7	3	10	7	4	0	-
	map00770	26	4	16	6	4	16	6	4	1	2
	map00780	11	5	8	2	2	8	2	2	2	2
	map00790	44	12	25	9	6	26	6	5	5	5
	map00670	9	2	5	3	3	4	1	6	1	2
	map00860	79	17	42	13	5	50	10	6	5	4
	map00130	41	6	23	11	5	25	9	4	1	2
	av	37.7	8.8	22.2	7.9	4.7	22.7	6.8	5.2	2.6	3.2

[a] The table shows the result of three types of analyses: the clustering of all compounds (all compounds), the pathway-oriented clustering (by pathway), and the matching of enzyme gene clusters and compound clusters (by EC), as well as the total number of compounds (CPD) and the total number of enzymes (EC) that are found in operons in certain genomes in the KEGG ortholog group tables. Num1 is the total number of clusters found, Num2 is the total number of clusters excluding singletons, Max is the number of members in the largest cluster, NumE is the number of enzyme gene clusters mapped onto pathways and containing at least one orthlog enzyme, and MaxC is the maximum number of chemical compounds in the intersection of the similar compound cluster and the enzyme gene cluster. Abbreviations for the pathway classes are: C, carbohydrate metabolism; E, energy metabolism; L, lipid metabolism; N, nucleotide metabolism; AA, amino acid metabolism; oAA, metabolism of other amino acids; CC, metabolism of complex carbohydrates; CL, metabolism of complex lipids; and CoV, metabolism of cofactors and vitamins.

(43) Parkhill, J.; Wren, B. W.; Thomson, N. R.; Titball, R. W.; Holden, M. T.; Prentice, M. B.; Sebaihia, M.; James, K. D.; Churcher, C.; Mungall, K. L.; Baker, S.; Basham, D.; Bentley, S. D.; Brooks, K.; Cerdeno-Tarraga, A. M.; Chillingworth, T.; Cronin, A.; Davies, R. M.; Davis, P.; Dougan, G.; Feltwell, T.; Hamlin, N.; Holroyd, S.; Jagels, K.; Karlyshev, A. V.; Leather, S.; Moule, S.; Oyston, P. C.; Quail, M.; Rutherford, K.; Simmonds, M.; Skelton, J.; Stevens, K.; Whitehead, S.; Barrell, B. G. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **2001**, *413*, 523−527.

(44) DelVecchio, V. G.; Kapatral, V.; Redkar, R. J.; Patra, G.; Mujer, C.; Los, T.; Ivanova, N.; Anderson, I.; Bhattacharyya, A.; Lykidis, A.; Reznik, G.; Jablonski, L.; Larsen, N.; D'Souza, M.; Bernal, A.; Mazur, M.; Goltsman, E.; Selkov, E.; Elzer, P. H.; Hagius, S.; O'Callaghan, D.; Letesson, J. J.; Haselkorn, R.; Kyrpides, N.; Overbeek, R. The genome sequence of the facultative intracellular pathogen *Brucella melitensis*. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 443−448.

(45) Kunst, F.; Ogasawara, N.; Moszer, I.; Albertini, A. M.; Alloni, G.; Azevedo, V.; Bertero, M. G.; Bessieres, P.; Bolotin, A.; Borchert, S.; Borriss, R.; Boursier, L.; Brans, A.; Braun, M.; Brignell, S. C.; Bron, S.; Brouillet, S.; Bruschi, C. V.; Caldwell, B.; Capuano, V.; Carter, N. M.; Choi, S. K.; Codani, J. J.; Connerton, I. F.; Cummings, N. J.; Daniel, R. A.; Denizot, F.; Devine, K. M.; Dusterhoft, A.; Ehrlich, S. D.; Emmerson, P. T.; Entian, K. D.; Errington, J.; Fabret, C.; Ferrari, E.; Foulger, D.; Fritz, C.; Fujita, M.; Fujita, Y.; Fuma, S.; Galizzi, A.; Galleron, N.; Ghim, S. Y.; Glaser, P.; Goffeau, A.; Golightly, E. J.; Grandi, G.; Guiseppi, G.; Guy, B. J.; Haga, K.; Haiech, J.; Harwood: C. R.; Henaut, A.; Hilbert, H.; Holsappel, S.; Hosono, S.; Hullo, M. F.; Itaya, M.; Jones, L.; Joris, B.; Karamata, D.; Kasahara, Y.; Klaerr-Blanchard, M.; Klein, C.; Kobayashi, Y.; Koetter, P.; Koningstein, G.; Krogh, S.; Kumano, M.; Kurita, K.; Lapidus, A.; Lardinois, S.; Lauber, J.; Lazarevic, V.; Lee, S. M.; Levine, A.; Liu, H.; Masuda, S.; Mauel, C.; Medigue, C.; Medina, N.; Mellado, R. P.; Mizuno, M.; Moestl, D.; Nakai, S.; Noback, M.; Noone, D.; O'Reilly, M.; Ogawa, K.; Ogiwara, A.; Oudega, B.; Park, S. H.; Parro, V.; Pohl, T. M.; Poetetelle, D.; Porwollik, S.; Prescott, A. M.; Presecan, E.; Pujic, P.; Purnelle, B.; Rapoport, G.; Rey, M.; Reynolds, S.; Rieger, M.; Rivolta, C.; Rocha, E.; Roche, B.; Rose, M.; Sadaie, Y.; Sato, T.; Scanlan, E.; Schleich, S.; Schroeter, R.; Scoffone, F.; Sekiguchi, J.; Sekowska, A.; Seror, S. J.; Serror, P.; Shin, B. S.; Soldo, B.; Sorokin, A.; Tacconi, E.; Takagi, T.; Takahashi, H.; Takemaru, K.; Takeuchi, M.; Tamakoshi, A.; Tanaka, T.; Terpstra, P.; Tognoni, A.; Tosato, V.; Uchiyama, S.; Vandenbol, M.; Vannier, F.; Vassarotti, A.; Viari, A.; Wambutt, R.; Wedler, E.; Wedler, H.; Weitzenegger, T.; Winters, P.; Wipat, A.; Yamamoto, H.; Yamane, K.; Yasumoto, K.; Yata, K.; Yoshida, K.; Yoshikawa, H. F.; Zumstein, E.; Yoshikawa, H.; Danchin, A. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **1997**, *390*, 249−256.

(46) Blattner, F. R.; Plunkett, G., 3rd; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y. The complete genome sequence of *Escherichia coli* K-12. *Science* **1997**, *277*, 1453−1474.

(47) Parkhill, J.; Dougan, G.; James, K. D.; Thomson, N. R.; Pickard, D.; Wain, J.; Churcher, C.; Mungall, K. L.; Bentley, S. D.; Holden, M. T.; Sebaihia, M.; Baker, S.; Basham, D.; Brooks, K.; Chillingworth, T.; Connerton, P.; Cronin, A.; Davis, P.; Davies, R. M.; Dowd, L.; White, N.; Farrar, J.; Feltwell, T.; Hamlin, N.; Haque, A.; Hien, T. T.; Holroyd, S.; Jagels, K.; Krogh, A.; Larsen, T. S.; Leather, S.; Moule, S.; O'Gaora, P.; Parry, C.; Quail, M.; Rutherford, K.; Simmonds, M.; Skelton, J.; Stevens, K.; Whitehead, S.; Barrell, B. G. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **2001**, *413*, 848−852.

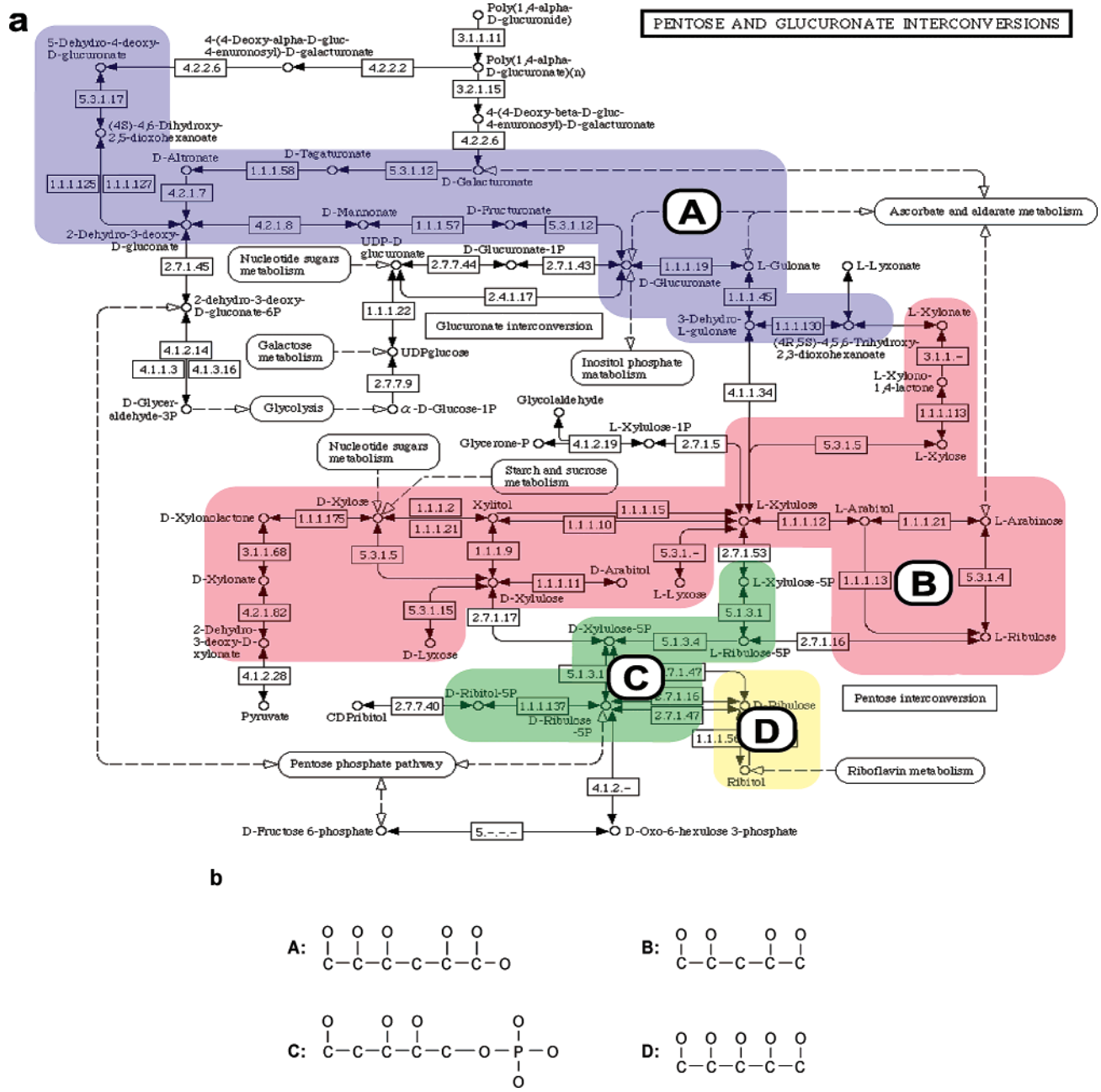

Figure 7. An example of similar compound clusters mapped onto a specific pathway. a is the result of the pathway-oriented clustering for the metabolic pathway of pentose and glucuronate interconversions, whose accession number is map00040 in the KEGG/PATHWAY database. After clustering 2294 metabolites that appear on any of the KEGG pathway maps, non-singleton clusters were superimposed on each of the pathway maps. Here, chemical compounds included in the same shaded region exhibit high structural similarities and high connectivities along the pathway in map00040. There are four major clusters of such chemical compounds in this pathway map: **A**, **B**, **C**, and **D** whose schematic representations of common components are drawn in b.

pathway. Here we have shown that the correlation exists between the structural similarity and the pathway connectivity of chemical compounds, and furthermore that the genomic/pathway correlation of enzymes and the chemical/pathway correlation of compounds do sometimes overlap. These two observations are best illustrated in the KEGG pathway map for pentose and glucuronate interconversions (http://www.genome.ad.jp/kegg/pathway/map/map00040.html).

First, this pathway map could be divided into two large clusters A and B (Figure 7) according to the structural similarity of chemical compounds. The difference of these two clusters is characterized by the difference of the number of carbon atoms; A is the glucuronate-related group and B is associated with pentoses. In fact, enzymatic reactions corresponding to the connector between two sub-pathways are lyases acting on carbons, such as a decarboxylase for reducing or raising the number of carbon atoms. Thus, we could identify biochemically meaningful clusters simply by comparison of chemical structures.

Second, there are at least six operon-like structures for the enzyme genes according to the KEGG ortholog group table that summarizes genomic contexts of completely sequenced genomes (see http://www.genome.ad.jp/kegg/ortholog/tab00040.html and also http://www.genome.ad.jp/kegg/pathway/ot/ot00040.html). Three of them were found to be highly correlated with cluster A (Figure 8 and Table 3). To summarize our observations,

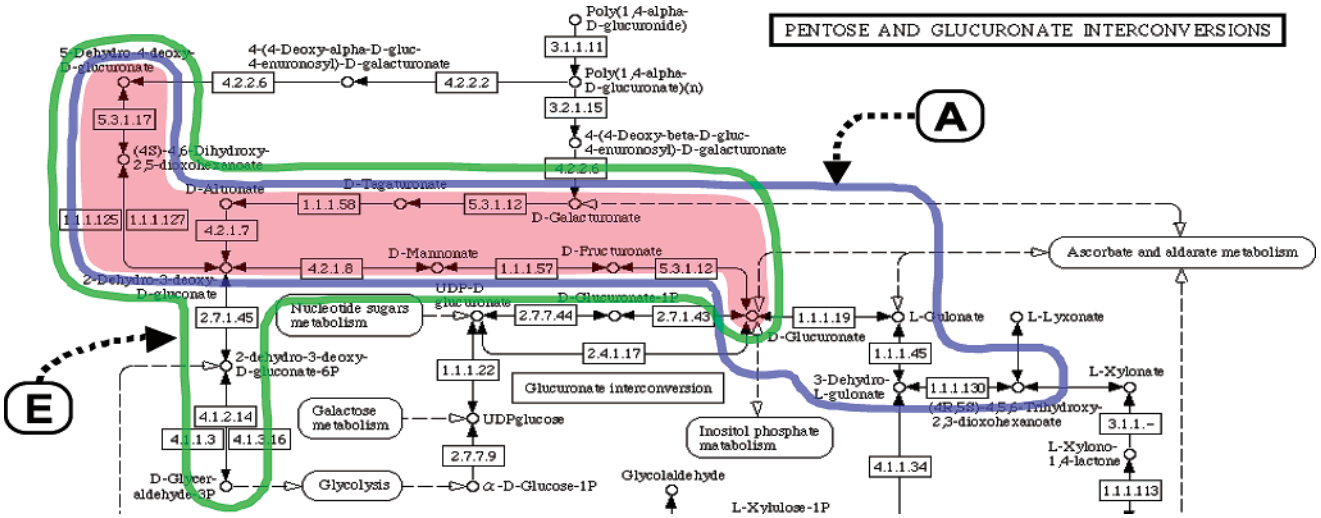

Figure 8. Example of the correlation between chemical information and genomic information. The area designated by A corresponds to the cluster of similar compounds shown in Figure 7. The area designated by E corresponds to the cluster of genomic associations where genes coding for the enzymes are closely located on selected genomes according to the KEGG ortholog group table. Thus, the shaded area is the overlap of chemical and genomic clusters.

Table 3. Overlap of Similar Compound Clusters and Enzyme Gene Clusters in the KEGG Pathway map00040

compound clusters[a]	enzyme gene clusters (possible operons)
A	4.2.1.7, 1.1.1.58, 5.3.1.12
A	4.2.1.8, 1.1.1.57, 5.3.1.12
A	5.3.1.17, 1.1.1.125, (2.7.1.45), (4.1.2.14), (4.1.3.16)
B, C	5.3.1.4, (2.7.1.16), 5.1.3.4, (2.7.1.53), (5.−.−.−), (4.1.2.−)
B	5.3.1.5, (2.7.1.17)
C, D	1.1.1.56, (2.7.1.47)

[a] Clusters of similar compounds A, B, C, and D correspond to those shown in Figure 7, and clusters of enzyme genes are taken from the KEGG ortholog group table. Each set of EC numbers in the same row represents a possible operon structure whose products are also adjacent on the metabolic pathway. The EC numbers in parentheses were outside of the overlap regions (see Figure 8).

chemical association may indicate pathway association, which in turn may indicate genomic association, and vice versa.

The KEGG metabolic pathway maps mostly represent intermediary metabolism, a core portion of the metabolic network that is shared and conserved in many different organisms. Among those maps, map00040 contained the largest intersection of chemical/pathway and genomic/pathway correlations. In other words, the intersection was smaller in the other KEGG maps. However, we expect to observe more examples of the three-way correlation of chemical/pathway/genomic clusters in secondary metabolism where environmental factors have more direct influences on genomic contents. Knowledge on chemical compounds can be utilized for gene annotations and pathway reconstructions in secondary metabolism where we have less knowledge on enzymes and more knowledge on chemical compounds. For example, special biosynthetic/biodegradation pathways in bacteria or special biosynthetic pathways in plants may be uncovered by analyzing structural similarities of chemical compounds and searching for clusters of possible enzyme genes in the genome.

The tendency that structurally similar compounds are closely positioned on the pathway can be confirmed by the distribution of compound similarity scores along the KEGG pathways (Figure 9). The average similarity score of compound pairs decreases as the distance of those pairs along the pathway increases, but there is a short-range correlation of similarity scores and pathway distances. This may reflect the nature of the metabolic pathways where each metabolite is modified little by little, thus forming clusters of similar compounds on the pathway maps.

Classification of Atomic Environments. In this study, chemical compounds were treated as 2D graph objects consisting of atoms (nodes) and atomic bonds (edges), namely, without considering 3D structures. However, to incorporate reactivity and other chemical properties that depend on three-dimensional aspects, compounds were viewed as consisting of functional groups, and the same atoms with different environments were distinguished accordingly. We took into account the group-contribution methods for estimating standard Gibbs energies of formation of biochemical compounds[28−30] when we defined the total of 68 atom types (Figure 1). The conversion from the MDL/MOL format to the 68-atom-type representation was done computationally[30] for all the KEGG compounds. Obviously, this is not the only way to classify atom types. In fact, we first defined about 90 atom types with finer classification of ring structures, but then the numbers of instances in the KEGG compound database were too small for some types. With the current classification we obtained reasonable results for comparison and clustering of KEGG compounds and for identification of common substructures. The usefulness of our classification should further be evaluated by different types of analyses (see below).

The atom type representation contains the information about not only the atom species but also neighbor atoms and bond patterns. Thus, it partially incorporates three-dimensional aspects of compounds. Although the current classification is not sufficient for distinguishing, for example, chilarity of compounds, such an additional feature may be included in a finer classification of atom types. Again, the validity of the finer classification should be examined by the usefulness of biochemical features detected. As for the atom types that are categorized into undefined classes in Figure 1e, they come from inorganic molecules or they have unusual bond structures such as R═C═R. The numbers of instances were too small to warrant consideration of separately defined environmental information.

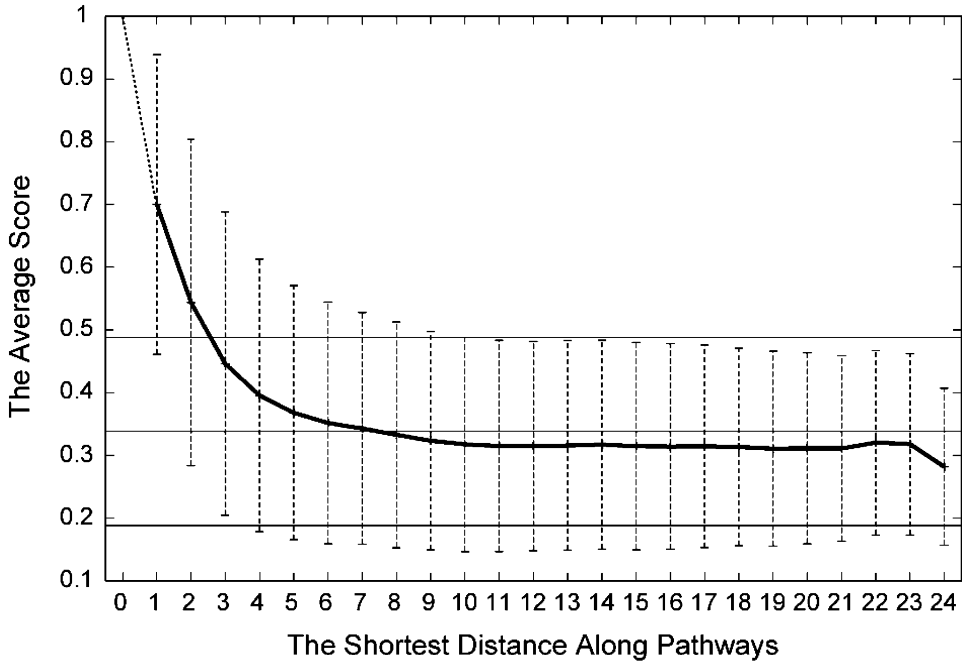

Figure 9. Average similarity score (thick line) and the standard deviation (dashed lines) are plotted against the distance for pairs of chemical compounds along the KEGG pathway. The distance is measured by the length of the shortest path along the pathway, which varies from 0 to 24. Here, the distance 0 means self-similarity, that is, the score is expected to be exactly 1. The average length of shortest paths was around 9 for all pairs along the pathways. The average similarity score μ and the standard deviation σ for all KEGG compounds (Figure 4) are also shown by the three horizontal lines corresponding to $\mu+\sigma$, μ and $\mu-\sigma$.

Similarity Measure for Compound Comparison. When comparing two chemical compounds, we used the three-value weighting scheme: 1 for a perfect match of atom types, 0.5 for a partial match of the same atomic species with different atom types, and 0 for a mismatch of atomic species. In principle, it should be possible to define a score matrix or a "mutation matrix" for all pairs of 68 atom types. For example, the scoring may be based on the 3D structural similarity of compounds. Alternatively, the scoring may be based on the reactivity between compounds or the closeness in terms of the chemical reaction steps, especially those catalyzed by enzymes. An appropriate measure of chemical reactivity should be useful not only for assessing closeness of compounds in biochemical pathways, but also for generating all possible compounds that can be converted from a given compound and predicting reaction pathways. Toward this end, we are experimenting a simple extension of the current three-value weighting scheme by distinguishing the matches of ring structures and chain structures. The classification of 68 atom types may also have to be reexamined from this perspective.

In some cases of the atom alignments generated by the SIMCOMP program, certain atoms that should be aligned were not included in the common substructure. First of all, when the relationship between two compounds was very distant, the conserved region was too small and the program misidentified the common substructure. Second, the association graph method was sometimes not effective, because the maximal clique found was not necessarily the best match but the best set of matches in the biochemical sense. These problems should be alleviated by introducing more appropriate weighting schemes. In the present analysis, however, the effect of such computational errors is negligible because our result of comparing chemical similarity with pathway and genomic information is based only on high scoring pairs.

Common Subgraphs and Cliques. The general problem of finding the maximal common subgraph of two graphs or finding the maximal clique is known to be NP-hard. However, our particular problem of comparing two chemical structures is not really NP-hard, because there is a clear limit for the number of edges at each node, i.e., the maximum of four for a carbon atom. The association graph method that we used is a general method for finding common subgraphs and we did not directly take into account this special graph structure. Although it may be feasible to develop a drastically different algorithm, the heuristics introduced in the traditional association graph method was sufficiently effective to identify biochemical features. We discontinued the clique finding procedure at a given number of steps and then looked for a better solution for each of the connected components (SCCSs) larger than a given size. Thus, these heuristics reduced the execution time and identified local matches, which we hoped were likely to be biochemically meaningful substructures.

To examine if this is in fact the case, we performed a comparison of the exact (optimal) solution and the heuristic (suboptimal) solution. Here a "virtually" exact solution was obtained by setting the maximum number of recursion steps R_{max} at a sufficiently large value (one million to 10 million). The heuristic solution was obtained as described; with $R_{max} =$ 15 000 and by searching for optimal SCCSs. We prepared two data sets of compound pairs: one randomly selected from the entire database of 9383 compounds, and the other taken from the neighboring pairs along the KEGG metabolic pathways, namely those having substrate-product relations in enzymatic reactions. As shown in Table 4, the performance of our heuristics was measured by the ratio m_h/m_e, the number of matched atoms m_h in the heuristic solution divided by the number of matched atoms m_e in the exact solution. The result indicates that although the heuristic method may fail to detect exact solutions in about

Table 4. Comparison of the Heuristic Algorithm with the Exact Algorithm

ratio[a] of matching, m_h/m_e	random pairs	pairs along pathways
equal to 1.0	157	185
equal to 0.8 − less than 1.0	37	12
less than 0.8	6	3

[a] Here, m_h is the size of atom matching by our heuristic algorithm, and m_e is that by the exact algorithm. Two data sets, each containing 200 compound pairs, are generated from the entire database (random pairs) and from the neighboring pairs along the metabolic pathways (pairs along pathways).

20% of randomly selected compound structure comparisons, it becomes more successful, with the missing rate of less than 10%, for the comparison of biochemically related compounds. By considering the 100 times faster computation time in Table 4, our heuristic method should be sufficient for detecting biochemically meaningful features.

However, we also noticed that an improvement was desirable for the choice of threshold parameters. We used the same number of steps to suspend the clique finding and the same cutoff to eliminate small SCCSs for all calculations. Because the sizes of the search space and the candidate set of quasi-MCSs are dependent on the compounds to be compared, it would be more effective to use proper parameters for each calculation. To estimate such parameter sets, we need to learn more about statistics of graph similarities and investigate biochemical results obtained with different parameters.

Availability. Each program in the SIMCOMP package is written in C language or Perl script language and intended to work well on most standard UNIX operating systems. All source codes are available from our web site http://web.kuicr.kyoto-u.ac.jp/simcomp/. One can find hardware and software requirements and detailed instructions for installation of the package.

Acknowledgment. We thank Dr. Tatsuya Akutsu for helpful discussions on the graph isomorphism problem of chemical compounds and Koichiro Tonomura, Rumiko Yamamoto, Tomoko Komeno, and Masaaki Kotera for checking the compound and reaction data in the course of preparing our dataset. We also thank all of the KEGG project team members for maintaining and updating the LIGAND and PATHWAY databases, without which this work would not have been possible. This work was supported by the 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. All of the computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Supporting Information Available: The database file of chemical compounds compiled from KEGG/LIGAND, list of correspondence between atom labels used in the above database file and KEGG atoms used in the manuscript, cluster tree of all compounds, and the experimental results of pathway and ortholog oriented clustering (text). This material is available free of charge via Internet at http://pubs.acs.org.

JA036030U