

## Extraction and Analysis of Chemical Modification Patterns in Drug Development

Daichi Shigemizu,<sup>†</sup> Michihiro Araki,<sup>‡</sup> Shujiro Okuda,<sup>†</sup> Susumu Goto,<sup>†</sup> and Minoru Kanehisa<sup>\*,†,‡</sup>

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan, and Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai Minato-ku, Tokyo 108-8639, Japan

Received October 14, 2008

Most drugs have been continuously modified from prototypic compounds in the drug development process. Such chemical modifications in the history of drug development are expected to contain a wealth of medicinal chemists' knowledge, and the KEGG DRUG structure maps have been compiled to capture this knowledge. Here we attempted to extract the information on the chemical modification patterns from 3745 approved drugs in the KEGG DRUG database and 255 drug pairs in the KEGG DRUG structure maps. We first identified 236 core structures and 506 peripheral fragments from the KEGG DRUG database using bit-represented fingerprints and hierarchical clustering of similar structures. We then examined position-dependent relationships between core structures and peripheral fragments, which revealed the tendency of specific fragments connected to specific modification sites on the core structures. Next we converted the drug pairs into 204 peripheral fragment changes at the modification sites. Each change was represented by the transformation profile defined as a difference of fingerprint bit patterns, and the hierarchical clustering of similar transformation profiles was performed. We thus identified 125 chemical modification patterns that characterize the KEGG DRUG structure maps. These patterns were further applied to the reconstruction of a new structure map. The approach presented here may be applicable to systematic *in silico* drug modifications.

### INTRODUCTION

Currently available drugs have been mostly derived from prototypic compounds found through empirical screening.<sup>1,2</sup> In the history of drug development, medicinal chemists have continuously introduced modifications around core chemical structures found in the prototypic compounds to improve efficacy or to apply to different therapeutic categories. Chemical and biological functions of drugs highly depend on the combinations of limited numbers of conserved core structures and diverse fragments around them. From this perspective, it would be useful to collect and computerize the collective knowledge of medicinal chemists on the chemical modifications in the drug development process and to extract distinct patterns for further bioinformatic analysis of drug structures. As part of the KEGG project, we have developed the KEGG DRUG resource and, in particular, the KEGG DRUG structure maps in order to capture this knowledge.<sup>3</sup> Here we report a method for extracting chemical modification patterns from the structure maps and chemoinformatic analysis using these patterns.

The KEGG DRUG structure maps (<http://www.genome.jp/kegg/pathway.html#drug>) illustrate structural pathways, showing the sequence of chemical modifications introduced to prototypic compounds during drug development. The structure maps are classified into three categories: chronology, target based structure classification, and skeleton based structure classification. Currently the structure maps in the chronology category are further classified into four subcategories: antibiotics, antineoplastics, nervous system agents,

and other drugs. We used the structure maps in the chronology category and manually extracted pairs of drugs connected by arrows in order to extract chemical modifications.

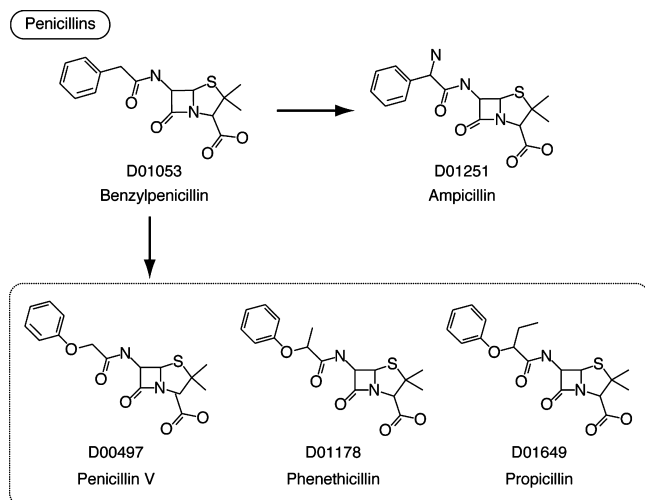
Since most drugs have been derived from prototypic compounds, they should be divisible into conserved core structures and peripheral fragments. In previous research, several computational tools have been developed for searching substructures, such as the automated ring searching<sup>4,5</sup> and shape descriptor methods.<sup>6–8</sup> For instance, Bemis and Murcko identified representative core structures by dividing each drug in the Comprehensive Medicinal Chemistry (CMC) database into ring, linker, framework, and side chain atoms using the shape descriptor method.<sup>9</sup> These methods have helped to clarify the combinatorial rules between core structures and peripheral fragments in current drugs, but little attention has been paid to understanding chemical modifications in terms of each core structure.

Here we used the KEGG DRUG database, which contains all approved drugs in both the USA and Japan, and performed hierarchical clustering of chemical structures to identify a set of core structures in the form of rings and linkers as well as a set of peripheral fragments after removing the core structure in each drug structure. We then used the KEGG DRUG structure maps, which represent a small but representative subset of drugs, and extracted the changes of peripheral fragments at specific modification sites on the core structures. The changes were characterized by a set of chemical modification patterns that depended on the core structures and the modification sites. The patterns were subsequently used to generate possible conversions of known drug structures. The strategy of computerizing such chemical modification

\* Corresponding author: [kanehisa@kuicr.kyoto-u.ac.jp](mailto:kanehisa@kuicr.kyoto-u.ac.jp).

<sup>†</sup> Kyoto University.

<sup>‡</sup> University of Tokyo.



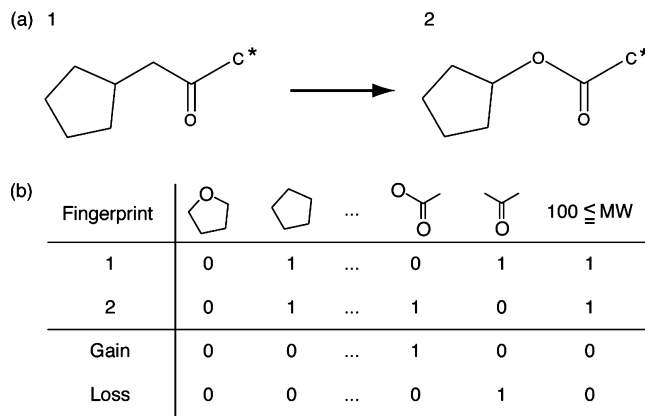
**Figure 1.** A part of the penicillins map in the KEGG DRUG structure map, where each drug is identified by the D number. An arrow shows the chemical modification from one product to another product or another group of products. Two drugs before and after the chemical modification constitute a drug pair. In this case four drug pairs are identified: (D01053 and D01251), (D01053 and D00497), (D01053 and D01178), and (D01053 and D01649).

patterns will be useful for understanding druglikeness and druglike modifications as well as for searching new modifications in existing drugs or natural compounds.

## MATERIALS AND METHODS

**Data Sets of Drugs and Drug Pairs.** All drug data used in this study were collected from the KEGG DRUG database,<sup>3,10</sup> containing 5417 approved drugs in both the USA and Japan as of March 2008. Each entry contains a D number (accession number), generic and trade names, formula, chemical structure, target information, activity information, therapeutic category linked to the KEGG BRITE database, and links to other databases. Nondrug data were extracted from the KEGG COMPOUND database. The chemical structure information (atoms and bonds) is available in the form of the MDL/MOL file. In the present analysis, we used 3745 drug structures excluding those entries for single atoms, mixtures, and supplements in KEGG DRUG. In order to analyze chemical modifications, we used the KEGG DRUG structure maps. As illustrated in Figure 1, the binary associations linked with arrows in the KEGG DRUG structure maps were manually defined as drug pairs.

**Bit-Represented Fingerprints for Chemical Structures.** To define similarities between two chemical structures, we introduced a bit-represented fingerprint aggregating 22 keys for physicochemical properties and 618 keys for substructural properties. The physicochemical properties were defined for the octanol/water coefficient or the logP value (8 keys), the number of rotatable bonds (8 keys), and the molecular weight (6 keys), which have often been used in evaluating drug-likeness.<sup>11</sup> The structural properties were selected from the dictionary-based fingerprint in PubChem (<http://pubchem.ncbi.nlm.nih.gov/>) excluding its section 1 (hierarchical element counts) and section 2 (rings in a canonic ESSSR ring set) (see Supporting Information S1). The PubChem fingerprint is composed of 881 substructure-keys (skeys) in total, which are similar in nature to the well-known MDL MACCS keys



**Figure 2.** (a) The chemical transformation of a drug pair is defined by the change of the peripheral fragments following removal of the common core structure. Atoms with asterisks indicate the modification site on the core structure. (b) The transformation profile is a difference of the bit-represented fingerprints for the two peripheral fragments, defined as a concatenation of the gain and loss profiles.

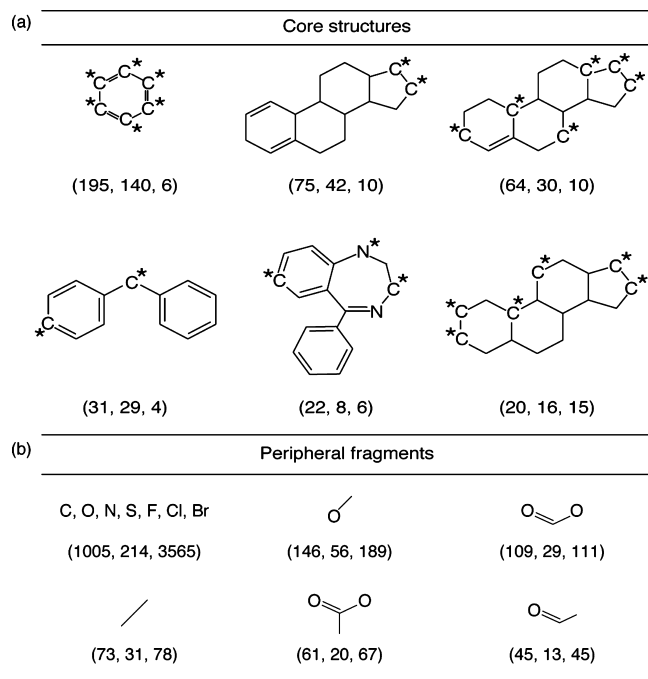
fingerprints. We converted each chemical structure into the bit-string representation, where each bit accounted for the presence or absence of the specific molecular feature.

**Comparison of Chemical Structures and Extraction of Common Substructures.** The similarity score between bit-string represented fingerprints for chemical structures was calculated using the Tanimoto coefficient.<sup>12</sup> A hierarchical clustering with UPGMA was performed using the distance matrix obtained from the similarity scores. The threshold of the similarity scores was set to 0.9 to define clusters for extracting common substructure. The common substructures were extracted in the form of the ring systems, which are cycles in the graph representation of chemical structures such as benzene, and the linker atoms directly connecting two ring systems.<sup>9</sup> We sometimes separated one computationally generated cluster into two or more clusters manually in order to distinguish heteroatoms and bond numbers. A set of extracted common substructures is considered as a set of core structures that make up chemical structures of drugs.

**Extraction of Chemical Modification Patterns.** Peripheral fragments are obtained after removing core structures and merging into similar groups. Drug pairs that represent chemical modifications were converted into transformation profiles of peripheral fragments. Each transformation profile is a difference of the bit-represented fingerprints for the two peripheral fragments. In practice it is a concatenation of the gain profile and the loss profile. The gain profile was defined as a bit-string focusing on the gain or neutral molecular bits after chemical modification, whereas the loss profile was defined as a bit-string focusing on the loss or neutral molecular bits after chemical modification. The similarity scores for the transformation profiles were also defined by the Tanimoto coefficient. Figure 2 illustrates how to convert a peripheral fragment change into a transformation profile.

## RESULTS

**Identification of Core Structures and Peripheral Fragments.** In order to identify a set of core structures from 3745 drugs in the KEGG DRUG database, we performed hierarchical clustering of chemical structures based on bit-string based similarity scores. We obtained 449 clusters



**Figure 3.** Frequently observed core structures and peripheral fragments in the KEGG DRUG database. (a) For the core structures, the numbers in parentheses indicate the number of drugs with the core structure, the number of peripheral fragments connected to the core structure, and the number of the modification sites on the core structure. The atoms with asterisks indicate the modification sites connected to at least two different peripheral fragments. (b) For the peripheral fragments, the numbers in parentheses indicate the number of drugs with the peripheral fragment, the number of core structures connected to the peripheral fragment, and the total number of occurrences of the peripheral fragment.

excluding singletons when the Tanimoto coefficient 0.9 was used as a threshold. The core structures were identified as common substructures consisting of rings and linkers, which were manually defined for 410 clusters after excluding 39 clusters composed of chains only. As 15 clusters out of the 410 clusters included similar chemical structures with and without heteroatoms, we defined multiple common substructures for such clusters. After merging identical substructures, 236 core structures were identified (see Supporting Information S2). The core structures corresponded to most of the core structures that Bemis and Murcko previously identified in the Comprehensive Medicinal Chemistry (CMC) database.<sup>9</sup>

Next we performed a similar analysis for the peripheral fragments. We identified 635 unique peripheral fragment structures after removing the core structure from each drug structure in the cluster. The removal was done by using the SUBCOMP chemical structure comparison program ([http://www.genome.jp/ligand-bin/search\\_compound](http://www.genome.jp/ligand-bin/search_compound)). We then applied the same hierarchical clustering procedure and obtained 506 peripheral fragment clusters, which we simply call peripheral fragments in this paper.

The frequency distribution of core structures and peripheral fragments follows a power law distribution in each case. Figure 3 shows the top six core structures and the top six peripheral fragments that are frequently observed in the KEGG DRUG database. Since we considered distinct heteroatoms and bond numbers when defining core structures, topologically identical core structures were observed. In

**Table 1.** Observed Frequency of Modification Sites Classified According to the Number of Different Peripheral Fragments That Were Attached to the Same Site

no. of modification sites	observed frequency				ratio $\leq 2$
	0	1	2	$\geq 3$	
1	1	7	-	-	100%
2	7	19	2	-	100%
3	11	20	5	1	97%
4	16	34	10	3	95%
5	11	19	4	1	97%
6	10	2	3	4	80%
7	5	2	2	0	100%
8	0	0	2	1	67%
9	0	1	2	1	75%
$\geq 10$	2	9	4	6	71%

**Table 2.** Distributions of Atom Types at the Single-Modification Sites and the Multimodification Sites

atom type	single-modification sites	multiplication sites
carbon on ring	86.3%	79.3%
nitrogen on ring	4.8%	13.3%
carbon on linker	7.0%	6.6%
nitrogen on linker	0.4%	0.4%
other atoms on linker	1.5%	0.4%
total	100.0%	100.0%

Figure 3(a) there are three core structures that are similar to steroid structures, but the modification sites marked by asterisks are different from each other. This way, our definition of core structures would better distinguish relationships to peripheral fragments and chemical modifications patterns.

**Relationships between Core Structures and Peripheral Fragments.** Drug structures in the KEGG DRUG database were found to be combinations of 236 conserved core structures and 506 peripheral fragments. Although the number of all possible combinations is large, it seems that the number of druglike combinations is relatively small, for specific core structures and specific peripheral fragments appear to be linked at specific modification sites. We thus examined the number of peripheral fragments linked at each modification site of each core structure. There were modification sites connected to only a single peripheral fragment (single-modification sites) and those connected to at least two peripheral fragments (multimodification sites). Table 1 shows the observed frequency of single- and multimodification sites against the number of all modification sites in the core structure. The result clearly indicates that the number of different fragments that can be attached is two or less for the majority of the modification sites, irrespective of the total number of modification sites in the core structure. In other words, most modification sites were involved in linking to specific fragments, and we focus on such modification sites with respect to the core structure.

We also examined whether there were differences in the atom types between the single-modification sites and the multimodification sites. The atoms at the modification sites could be divided into five types: carbon on ring, nitrogen on ring, carbon on linker, nitrogen on linker, and the other atoms on linker. Table 2 shows the distributions of atom types at 892 single-modification sites and 256 multimodifi-



**Table 3.** Statistics of the Drug Pairs Extracted from the KEGG DRUG Structure Maps

number of drug pairs	255
number of drugs (and compounds)	269 (17)
number of modification sites	360
number of peripheral fragment changes	204
number of peripheral fragments	179

cation sites that were obtained from the 236 core structures. As can be seen, the distribution of atoms on linker was similar, whereas the distribution of atoms on ring was more different. The proportion of nitrogen atom on ring was significantly higher in the multimodification sites than in the single-modification sites.

Finally, we examined the boundary atoms on the peripheral fragments that were linked to each multimodification site. We found that 176 (69%) out of 256 multimodification sites were conserved in terms of the same boundary atoms on the peripheral fragments. These results suggest that the connection patterns between core structures and peripheral fragments tend to be limited even at the multimodification sites.

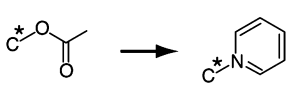
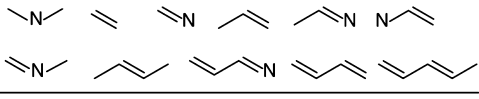
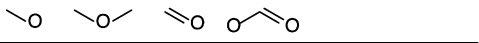
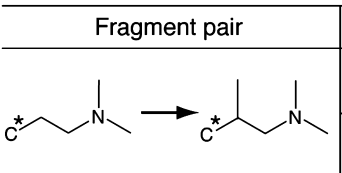

**Statistics of Drug Pairs.** The comprehensive analysis using the drug data from the KEGG DRUG database suggested that most drugs are composed of a limited number of combinations involving core structures and peripheral fragments. In addition, the peripheral fragments are found to be attached to frequently replaced sites on the core structures. This indicates that chemical modifications have been continuously introduced at specific sites on the core structures by medicinal chemists in the history of drug development. KEGG DRUG structure maps have been compiled to capture such medicinal chemists' knowledge, from which pairwise relationships of drug structures can be extracted. Here, we focused on 28 structure maps in the chronology category and manually defined 255 drug pairs consisting of 269 drugs and 17 nondrugs (Table 3). Since 76 out of 255 drug pairs contained chemical modifications at multiple sites, a total of 360 modification sites were identified. By using the set of 236 core structures, each drug pair was divided into a common core structure and a pair of peripheral fragments. We collected the changes of peripheral fragments at 360 modification sites and identified 204 unique changes consisting of 179 peripheral fragments.

**Characterization of Chemical Modifications.** In order to characterize the chemical modifications in drug pairs, we represented 204 peripheral fragment changes by the transformation profile, a bit string composed of a gain profile and a loss profile (Figure 2) and performed the hierarchical clustering of similar transformation profiles. We obtained 125 clusters using the threshold similarity score of 0.6. For each cluster the extraction of common bits was performed, and 125 chemical modification patterns were identified. Examples of the chemical modification patterns are shown in Figure 4, where the modification sites on the core structures are also included and marked by asterisks. The chemical modification from carboxyl group to nitrobenzene was represented by 11 gain bits and four loss bits (Figure 4a). The chemical extension of a methyl group was represented by one gain bit (Figure 4b). All the chemical modification patterns are shown in Supporting Information S3.

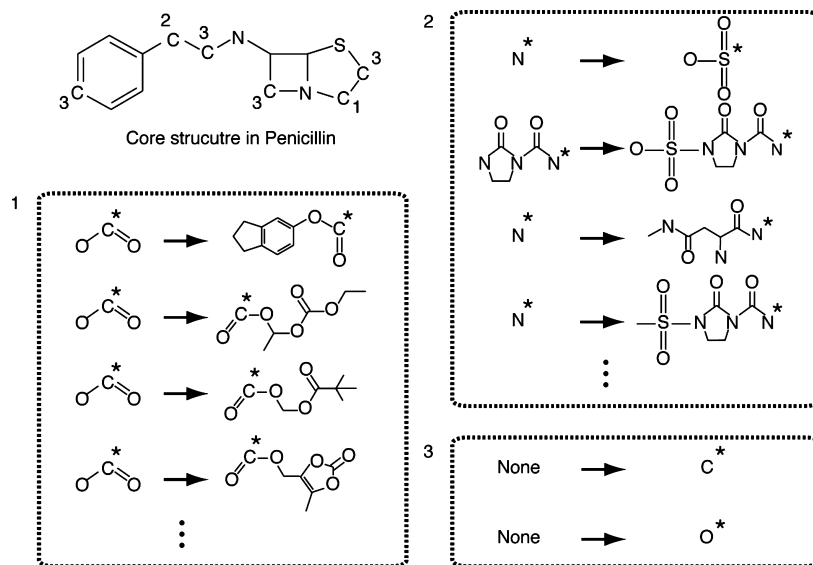
The chemical modification patterns were further classified into core-dependent chemical modification patterns that appeared only in specific core structures and core-independent chemical modification patterns that appeared in at least two different core structures. The numbers of core-dependent and core-independent chemical modification patterns were 92 (74%) and 33 (26%), respectively. In terms of the number of peripheral fragment changes, they corresponded to 117 (57%) and 87 (43%), respectively. In fact, the examples shown in Figure 4a and 4b represented a core-dependent chemical modification pattern and a core-independent chemical modification pattern. Although the fraction of core-independent patterns was one-fourth (26%) of all chemical modification patterns, these patterns were shared by many peripheral fragment changes as indicated by the larger fraction of actual numbers (43%).

**A Structural Framework of Chemical Modifications.** The identification and characterization of 125 chemical modification patterns may be used to systematically examine possible chemical structures for a given core structure. Figure 5 illustrates a structural framework of chemical modifications with respect to the core structure of the penicillin family. The core structure has six modification sites composed of four single-modification sites (marked by 3) and two multimodification sites (marked by 1 and 2). The applicable chemical modification patterns are shown in boxes with the corresponding numbering of modification sites. The chemical modification patterns in box 3 are applicable to all single-modification sites, while the chemical modification patterns in boxes 1 and 2 are applicable at the multimodification sites. In this case 7 out of 11 patterns in boxes 1 and 2 were core-independent chemical modification patterns commonly used in different core structures. The library of chemical modification patterns is thus linked to the core structures for systematic generation of possible drug structures.

**Reconstruction of the Structure Map.** We then examined if the library of chemical modification patterns could be applied to the reconstruction of drug developmental history as summarized in the KEGG DRUG structure maps. For this purpose, we considered all pairs of drugs in each group sharing one of the 236 core structures and collected only those pairs that matched to any of the chemical modification patterns in the library. As a result we obtained drug pair networks, each of which consisted of drugs with the same core structure. The two largest drug pair networks having steroid-like core structures shown in Figure 6(a) corresponded to the third and the sixth largest clusters in Figure 3(a). The first network consisted of 50 drugs, among which 42 drugs formed a connected subnetwork. (The corresponding cluster in Figure 3(a) contained 64 drugs because it included singletons without modification partners.) The second network consisted of 12 drugs, among which 6 drugs formed a connected subnetwork and none of them appeared in the KEGG drug structure maps. Thus, these 6 drugs were further examined in terms of the year introduced and its efficacy using the Medical Subject Headings (MeSH). As shown in Figure 6(b), the connection patterns of the drug pair network reproduced the history of drug development except the chronological order of D04300 and D05041. We conclude that the chemical modification patterns can provide sufficient information on constructing new structure maps in terms of the core structure.

(a)	Fragment pair	Chemical modification pattern
		<div data-bbox="699 184 754 212">Gain</div> <div data-bbox="785 163 1262 262">  </div> <div data-bbox="699 268 754 296">Loss</div> <div data-bbox="785 262 1262 304">  </div>
(b)	<div data-bbox="346 373 686 546">  </div>	<div data-bbox="699 436 754 464">Gain</div> <div data-bbox="785 436 1262 493">  </div> <div data-bbox="699 506 754 533">Loss</div> <div data-bbox="785 506 1262 533">None</div>

**Figure 4.** Examples of (a) a core-dependent chemical modification pattern and (b) a core-independent chemical modification pattern. Features of gain and loss bits are graphically shown.



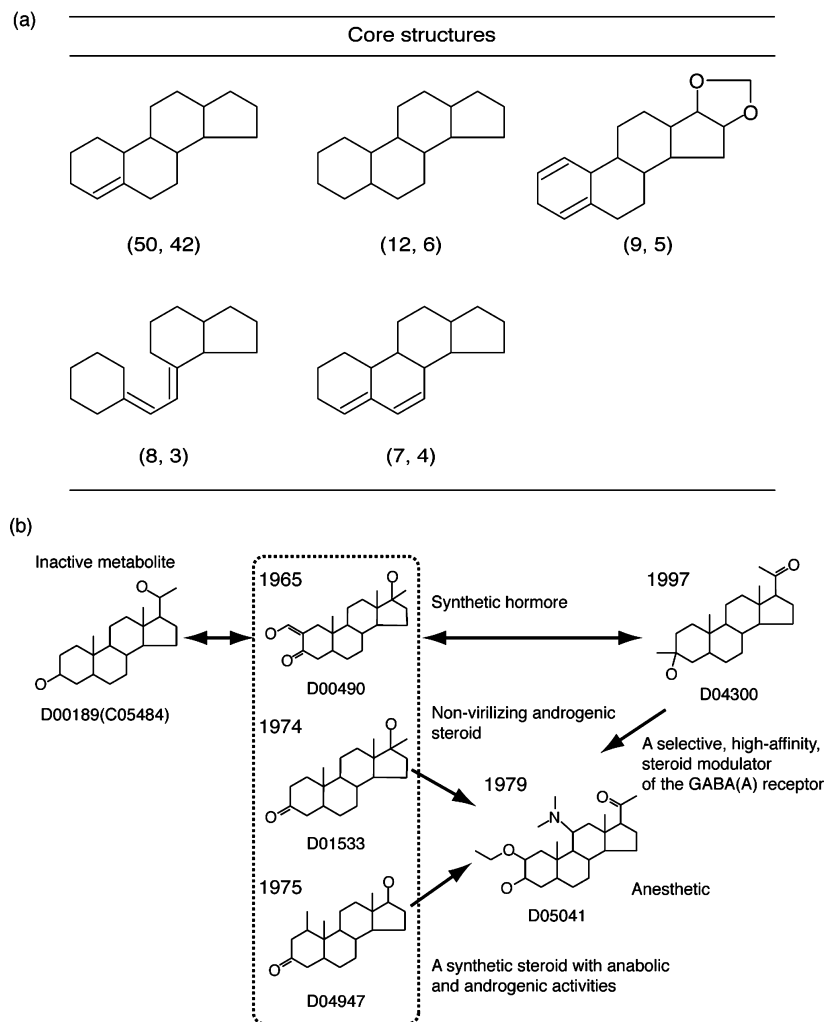
**Figure 5.** Systematic generation of penicillin-like compounds, where the penicillin core structure is modified according to the library of site-dependent chemical modification patterns. The numbering on the core structure indicates the distinction of modification sites and corresponding modification patterns (see text for details).

## DISCUSSION

Recent drug discovery efforts focus on making full use of high-throughput technology, refined computational tools, and accumulated knowledge to select candidates efficiently out of the huge chemical search space.<sup>13–16</sup> A number of *de novo* design programs and algorithms have been developed for proposing novel chemical structures.<sup>17–19</sup> The distributions of the chemical properties of drug and lead compounds have been analyzed to provide different kinds of chemical rules for druglikeness<sup>11,20–22</sup> and leadlikeness.<sup>23,24</sup> In contrast, our analysis is based on the chemical structures of approved drugs in the KEGG DRUG database and the history of drug development, such as from natural products to generations of marketed drugs, in the KEGG DRUG structure maps. Therefore, the chemical modification patterns that we identified in this analysis have more relevance on the optimization process than the discovery process. Indeed, the library of 125 chemical modification patterns contains well-known optimization processes, such as fluorine replacement on aromatic ring, which is used for metabolic site blocking,<sup>25,26</sup> and cyclization and ring size change, which are used for improving metabolic stability. The computerized

knowledge on such human-made chemical modification patterns may be used for finding new modifications of existing drugs, new uses of old drugs, and perhaps new drugs from natural products.

In this paper, we used bit representation of fingerprints, rather than graph representation, for 2D chemical structures in order to better identify chemical similarity of drugs and to define chemical transformation patterns. Fingerprints are often used in rapid similarity search tools, and many fingerprint-based search techniques have been developed including hashed connectivity pathways,<sup>27</sup> structural dictionary-based searches,<sup>28</sup> and layered atom environmental fingerprints.<sup>29</sup> In comparison to the rigorous definition and optimization in graph-based methods, fingerprints are defined and used in more subjective ways. It is difficult to accept any fingerprint as a standard for similarity search, but it is easy to include additional features, for example, other than structural properties. The purpose of drug structure similarity searches is not only simply the matching of chemical structures but also the identification of similar efficacy. As demonstrated by the Lipinski rule,<sup>11</sup> the importance of the physicochemical properties in druglikeness is well-known.



**Figure 6.** (a) The core structures for the five largest drug pair networks. The numbers in parentheses indicate the number of drugs having the core structure and the number of drugs that form the largest connected subnetwork. (b) A drug structure map reconstructed from the connected components of the second drug pair network. Arrows reflect the directions of the chemical modification patterns in the library.

Thus, we have developed our own fingerprint that incorporates the physicochemical properties, which are more global features of chemical compounds, in addition to the local substructural features. The core structures obtained by our fingerprint representation are compared with the core structures identified in previous research,<sup>9</sup> confirming that our fingerprint can be sufficiently applied to the similarity search of drug structures.

Another advantage of the fingerprint representation is that it is straightforward to measure the difference between two chemical structures, which can be used to define chemical transformation patterns for pairs of drugs. In fact, we previously used the graph based approach for the entire work presented in this paper. We first performed the hierarchical clustering of drug structures using the SIMCOMP program,<sup>30</sup> a graph based method to detect maximal common substructures, and obtained the core structures and peripheral fragments. We then defined substructure-based patterns to characterize chemical structure transformations in drug development, in a similar way to define the RDM patterns in biochemical reactions.<sup>31,32</sup> The result of the first part was very similar to the one reported here, but the chemical transformation patterns obtained by the graph-based method were too detailed and it was impossible to generalize the

patterns for use in further analysis. The fingerprint representation, especially when global features are incorporated, is much better suited for capturing the knowledge on chemical modification patterns in drug development. Of course, it is difficult to say how well the particular bit-string that we use now is optimized, but our method is general enough to incorporate potential improvements by adding/deleting specific features.

One possible direction for identifying druglike features is to examine different core structures with common chemical modifications. Since we converted all chemical structures into fingerprints, we could extract common properties from the bit-representation. We compared the core structures with common chemical modifications by focusing on the physicochemical bits. One common property found was the logP value or lipophilicity, and most of the core structures were similar to each other in this respect (see Supporting Information S4). The other physicochemical properties such as the molecular weight, H-bonding properties, and the number of rotatable bonds were distinctly different from each other. Optimal lipophilicity has been used in improving metabolic stability of compounds<sup>33</sup> and development of prodrugs.<sup>34</sup> Dragovich et al. demonstrated in a study of the human rhinovirus 3C protease inhibitor that the lead compound is

subjected to pharmacokinetic optimization for reducing lipophilicity without compromising activity.<sup>35</sup> Additionally, the prodrug approach has been applied to optimize lipophilicity and the desired durations of action.<sup>36</sup> Therefore, the chemical modifications extracted for the core structures seem to have preserved these important properties by keeping similar lipophilicity.

Based on a comprehensive analysis of the KEGG DRUG database we divided all drug structures into core structures and peripheral fragments. The division has revealed which sites on the core structures are involved in chemical modifications and which peripheral fragments are attached to which sites. We then used a limited data set of drug pairs taken from the KEGG DRUG structure maps to examine which peripheral fragments are modified to which other fragments at these sites. We have started collecting organic reactions including those used for drug development as part of the KEGG REACTION database. Thus, the drug pair data set will be expanded to better represent the lead optimization processes and, by combining with the existing reactant pair data set for enzymatic reactions, prodrug designs as well. We are also developing a software tool to predict druglike chemical modifications for a given compound by systematically generating possible fragment changes (see Figure 5) and somehow ranking them according to a measure of druglikeness.

An important extension of this study would be to incorporate information about target molecules as well as adverse reactions and interactions in the context of molecular networks such as those represented in the KEGG PATHWAY database. For example, frequently modified sites on drug structures may be related to selective binding sites for classes of target molecules. This way, we would have better knowledge on possible chemical and genomic spaces for drug-target interactions.

#### ACKNOWLEDGMENT

We thank Dr. Nobuya Tanaka for use of the ChemRuby library resource and Dr. Nelson Hayes for critical reading of our manuscript. This work was supported by the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency and the 21st Century COE Program "Genome Science" and a grant-in-aid for scientific research on the priority area "Comprehensive Genomics," both from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

**Supporting Information Available:** Bit-represented fingerprints used in this paper (Figure S1), representative core structures in the KEGG DRUG database (Figure S2), chemical modification patterns identified from drug pairs (Figure S3), and specific chemical modifications involved in lipophilicity (Figure S4). This material is available free of charge via the Internet at <http://pubs.acs.org>.

#### REFERENCES AND NOTES

- (1) Drews, J. Drug discovery: a historical perspective. *Science* **2000**, 287, 1960–1964.
- (2) Siegel, M. G.; Vieth, M. Drugs in other drugs: a new look at drugs as fragments. *Drug Discovery Today* **2007**, 12, 71–79.
- (3) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, 34, D354–357.
- (4) Domokos, L. Beilstein Ring Search System. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1993**, 33, 663–667.
- (5) Fan, B. T.; Panaye, A.; Doucet, J. P.; Barbu, A. Ring perception. A new algorithm for directly finding the smallest set of smallest rings from a connection table. *J. Chem. Inf. Comput. Sci.* **1993**, 33, 657–662.
- (6) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 82–85.
- (7) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, 27, 82–85.
- (8) Bemis, G. W.; Kuntz, I. D. A fast and efficient method for 2D and 3D molecular shape description. *J. Comput.-Aided Mol. Des.* **1992**, 6, 607–628.
- (9) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, 39, 2887–2893.
- (10) Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **2008**, 36, D480–484.
- (11) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, 46, 3–26.
- (12) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, 11, 1046–1053.
- (13) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, 1, 727–730.
- (14) Oprea, T. I. Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **2002**, 6, 384–389.
- (15) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature (London)* **2004**, 432, 855–861.
- (16) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, 24, 805–815.
- (17) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, 303, 1813–1818.
- (18) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, 4, 649–663.
- (19) Jonsdottir, S. O.; Jorgensen, F. S.; Brunak, S. Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics* **2005**, 21, 2145–2160.
- (20) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, 45, 2615–2623.
- (21) Feher, M.; Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 218–227.
- (22) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physicochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, 46, 1250–1256.
- (23) Sneader, W. *Drug Prototypes and their Exploitation*; Wiley: Chichester, New York, 1996.
- (24) Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 856–864.
- (25) Earl, J.; Kirkpatrick, P. Fresh from the pipeline. Ezetimibe. *Nat. Rev. Drug Discovery* **2003**, 2, 97–98.
- (26) Pui, C. H.; Jeha, S. Clofarabine. *Nat. Rev. Drug Discovery* **2005**, S12–13.
- (27) James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual*; Daylight Chemical Information Systems: Aliso Viejo, CA, 2008.
- (28) Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 141–142.
- (29) Bender, A. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1708–1718.
- (30) 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.
- (31) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* **2004**, 126, 16487–16498.
- (32) Oh, M.; Yamada, T.; Hattori, M.; Goto, S.; Kanehisa, M. Systematic analysis of enzyme-catalyzed reaction patterns and prediction of



- microbial biodegradation pathways. *J. Chem. Inf. Model.* **2007**, *47*, 1702–1712.
- (33) Nassar, A. E.; Kamel, A. M.; Clarimont, C. Improving the decision-making process in the structural modification of drug candidates: enhancing metabolic stability. *Drug Discovery Today* **2004**, *9*, 1020–1028.
- (34) Rautio, J.; Kumpulainen, H.; Heimbach, T.; Oliyai, R.; Oh, D.; Jarvinen, T.; Savolainen, J. Prodrugs: design and clinical applications. *Nat. Rev. Drug Discovery* **2008**, *7*, 255–270.
- (35) Dragovich, P. S.; Prins, T. J.; Zhou, R.; Johnson, T. O.; Hua, Y.; Luu, H. T.; Sakata, S. K.; Brown, E. L.; Maldonado, F. C.; Tuntland, T.; Lee, C. A.; Fuhrman, S. A.; Zalman, L. S.; Patick, A. K.; Matthews, D. A.; Wu, E. Y.; Guo, M.; Borer, B. C.; Nayyar, N. K.; Moran, T.; Chen, L.; Rejto, P. A.; Rose, P. W.; Guzman, M. C.; Dovalsantos, E. Z.; Lee, S.; McGee, K.; Mohajeri, M.; Liese, A.; Tao, J.; Kosa, M. B.; Liu, B.; Batugo, M. R.; Gleeson, J. P.; Wu, Z. P.; Liu, J.; Meador, J. W., 3rd; Ferre, R. A. Structure-based design, synthesis, and biological evaluation of irreversible human rhinovirus 3C protease inhibitors. 8. Pharmacological optimization of orally bioavailable 2-pyridone-containing peptidomimetics. *J. Med. Chem.* **2003**, *46*, 4572–4585.
- (36) Fredholt, K.; Larsen, D. H.; Larsen, C. Modification of in vitro drug release rate from oily parenteral depots using a formulation approach. *Eur. J. Pharm. Sci* **2000**, *11*, 231–237.

CI8003804