

Article

Toward Designing Drug-Like Libraries: A Novel Computational Approach for Prediction of Drug Feasibility of Compounds

Jing Wang, and Kal Ramnarayan

J. Comb. Chem., **1999**, 1 (6), 524-533 • DOI: 10.1021/cc990032m • Publication Date (Web): 19 October 1999

Downloaded from <http://pubs.acs.org> on March 20, 2009

More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 2 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

[View the Full Text HTML](#)



ACS Publications
High quality. High impact.

Toward Designing Drug-Like Libraries: A Novel Computational Approach for Prediction of Drug Feasibility of Compounds

Jing Wang* and Kal Ramnarayan

Structural Bioinformatics Inc., 10929 Technology Place, San Diego, California 92127

Received June 25, 1999

Prediction of the degree of drug-like character in small molecules is of great industrial interest. The major barrier, however, is the lack of a definition for drug-like character. We used the concept of the multilevel chemical compatibility (MLCC) between a compound and a drug library as a measure of the drug-like character of a compound. The rationale is that the local chemical environment of each atom or group of atoms in a compound largely contributes to the stability, toxicity, and metabolism *in vivo*. A systematic comparison of the local environments within a compound and those within the existing drugs provides a basis for determining whether and how much a compound is drug-like. We applied the MLCC calculations to four test sets: top selling drugs, compounds under biological testing prior to the preclinical test, anticancer drugs, and compounds known to have poor drug-like character. The following conclusions were obtained: (1) A convergent number of unique local structure types were found in the analysis of the library of the existing drugs. It suggests that the current drug library contains about 80% of all the viable types; therefore, discovery of a drug with new local structures is only an event of relatively small probability. (2) The method is highly selective in discerning drug-like compounds: most of the top drugs are predicted to be drug-like, about one-quarter of the biological testing compounds are drug-like, and about one-fifth of the anticancer drugs are drug-like. (3) The method also correctly predicted that none of the known problematic compounds are drug-like. (4) The method is fast enough for computational screening of virtual combinatorial chemistry libraries and databases of available compounds.

1. Introduction

The current technologies of drug discovery and development involve many stages, usually including initial screens, iterative lead optimizations, various pharmacokinetic tests, and final clinical trials. A strategy that allows focusing research efforts on the promising compounds in any of the stages would greatly increase the speed and cost-effectiveness of the entire process. Some groups have constructed drug-like combinatorial chemistry libraries¹ or selected drug-like compounds for screening.^{2,3} Thus, the initial lead discovery stages would generate promising candidates for later development. Continuing in this direction, we developed a computational strategy for assessing the potential of compounds to act as drugs. It can be applied to any stage where one desires focusing on drug-like candidates. In accordance to the principles that will be elucidated later, our method is named multilevel chemical compatibility (MLCC) calculation.

We considered that the chemical features of the substructures constituting a molecule are of primary importance in determining whether it would exhibit good *in vivo* behavior. For example, compounds bearing either reactive groups or bonds, which are easily broken with or without enzymes, are usually not suitable as drug candidates. Toxicity is often reported to relate to the existence of certain “toxicophore” groups, since these groups or their metabolites can undergo undesirable interactions with *in vivo* targets.^{4,5} Similarly, the

absorption, distribution, metabolism, and excretion of a drug are all dependent on the substructures which constitute the drug. Therefore, to determine whether a compound should be selected as a drug candidate, it is rational to compare its constituent substructures with those of existing drugs.

We further reasoned that not only the substructures but also the local chemical environments surrounding substructures are important for determining the *in vivo* behavior of a compound. For example, a structural unit such as a carbonyl group (C=O) may or may not be tolerated *in vivo* depending on the nature of its surrounding atoms. To reflect this fact, we developed the concept of “local structures” which describes the local chemical environment around each atom or group of atoms within a molecule. A compound is considered to bear drug-like characteristics if all its atoms or groups are situated in similar environments to those found in existing drugs. The substructures and local environments are examined at multiple levels of atom grouping.

The local structures of a compound were described by a set of n -centered groups. An n -centered group is defined as the part of the molecule formed by n contiguously connected non-hydrogen atoms plus the immediate neighboring atoms. The monocentered, dicentered, tricentered, and tetracentered groups were calculated for every entry in a drug library. The found groups were categorized into types. In examining a compound, the multicentered groups of the compound were similarly calculated and compared with the types from the drug library. A compound was considered as drug-compatible at level $n + 1$ if all the n -centered groups from the compound

* To whom correspondence should be addressed. Tel: (858) 675-2400. Fax: (858) 451-3828. E-mail: jing@strubix.com.

are consistent with the types from the drug library. The analysis was applied to a set of top selling drugs, compounds under biological testing, anticancer drugs, and known problematic molecules, respectively. It turned out that the calculations clearly reflect the feasibility of compounds to be used as drugs.

A potential problem with this approach is that the existing drugs may not have exhausted all the possible local structure types. Thus, the compounds with novel local motifs may artificially be ranked as non-drugs. We performed a convergence analysis for the number N of unique local structure types in the current drug library. We found that N approaches convergence when the contributing drugs increase. Most local structure types viable to construct drugs have been used in the existing drugs, and those unsuitable have been filtered out as a result of the experimental selections in drug discovery and development.

Few publications have addressed the questions about discriminating drugs and non-drugs. Sadowski and Kubinyi² used 120 atom types of Ghose and Crippen⁶ as descriptors of chemical structures and trained a neural network to discriminate drug molecules and non-drug molecules in terms of the constituent atom types. Ajay et al.³ used 166 ISIS substructure keys as descriptors and also a neural network as a discrimination learning tool. Other related publications include those of Bemis and Murcko⁷ and Lewell et al.,⁸ which extract information from existing drugs but are not designed for calculation of drug-like character. The common feature of these methods is that they use a set of predefined substructures as molecular descriptors. In contrast, the strength of our approach is that we use a general rule to describe any type of substructure. A large number of group types were identified from the drug library. For example, 949, 3995, 18748, and 69646 group types were identified for mono- to tetracentred groups, respectively. Such a large variety of group types cannot be covered by an empirically given, predefined set of chemical fragments. With fewer descriptors, the neural network technologies tend to recognize the potential patterns in composition of the descriptors within molecules as a basis for discrimination of drugs and non-drugs. The success of the pattern recognition depends on the constructs of the neural networks, definitions of the descriptors, and existence or not of any common patterns shared among drugs. In comparison, the MLCC calculation systematically searches for all the "local patterns", or local structures, which avoids the uncertainties associated with the neural network technologies.

MLCC calculation is conceptually different from neural network and similarity approaches. It determines whether a compound is drug-like based on compatibility and level of compatibility of the local structures, while disregarding overall similarity of a compound to drugs. This allows novel compounds to be ranked as drugs. In contrast, neural network approaches depend on the existence of certain collective, "recognizable" features shared among drugs. Similarity approaches are based on overall similarity of a compound and an existing drug. The shortage with the latter is that not only novel compounds cannot be ranked as drugs, but also compounds bearing reactive or toxic groups may be ranked

Table 1. Definition of Atom Types Used in the MLCC Calculations^a

type	definition
C2	sp ² carbon
C1	non-sp ² carbon
N2	sp ² nitrogen
N1	non-sp ² nitrogen
O2	sp ² oxygen
O1	non-sp ² oxygen
G1	any Cl, Br, or I

^a For all other cases, each element belongs to a different type.

as drugs due to the overwhelming similarity in the other parts of the compounds with some existing drugs.

2. Methodology

The constituents of a molecule can be examined at various levels of atom grouping. The lowest grouping would be the individual atoms contained in the molecule. A higher grouping is an atom plus its neighbor atoms. Even higher groupings can be obtained by extending the group sizes in a systematic way. The higher the grouping level at which the molecule is examined, the greater is the knowledge of chemical features. Using a systematic approach, we started from the lowest level to a grouping with four atom centers with neighbor atoms. In all the analyses, hydrogen atoms were omitted.

Atom Types—Level 1. An atom type is assigned to each heavy atom in a molecule. The definition of atom types is given in Table 1. The definition was kept as simple as possible since the environmental dependence of atoms is treated mainly by multilevel groupings, rather than by dividing atoms into types. However, we separated the atoms with sp² hybridization from those without sp² hybridization. Such a simple separation will aid in capturing the π -electron conjugation. The unification of chlorine, bromine, and iodine into one type is due solely to their similar chemical features (even though they have diverse biochemical properties), which are somewhat distinct from fluorine.

Monocentered Group—Level 2. Any heavy atom, together with its bonded heavy neighbors, defines a mono-centered group (Figure 1a). It is recorded by a character string such as

$$| C1 | 1 | A1 | 2 | A2 | 1 | A3 | \dots \quad (1)$$

where the vertical bar is used for ease of readability and should not be counted as a character. The first two characters are designated for indicating the atom type of the core atom. The third character designates the order of the bond between the core atom and its first neighbor. The fourth and fifth characters designate the type of first neighbor. The sixth is for the order of the bond core to second neighbor, and the seventh and eighth are for the type of second neighbor. It continues accordingly until all the immediate neighbors are described. The neighbor atoms should be arranged in a descending order according to their values. The value w of a neighbor atom with type such as A1 and bond order n is defined as

$$w = 'A' * 100000 + '1' * 10 + n \quad (2)$$

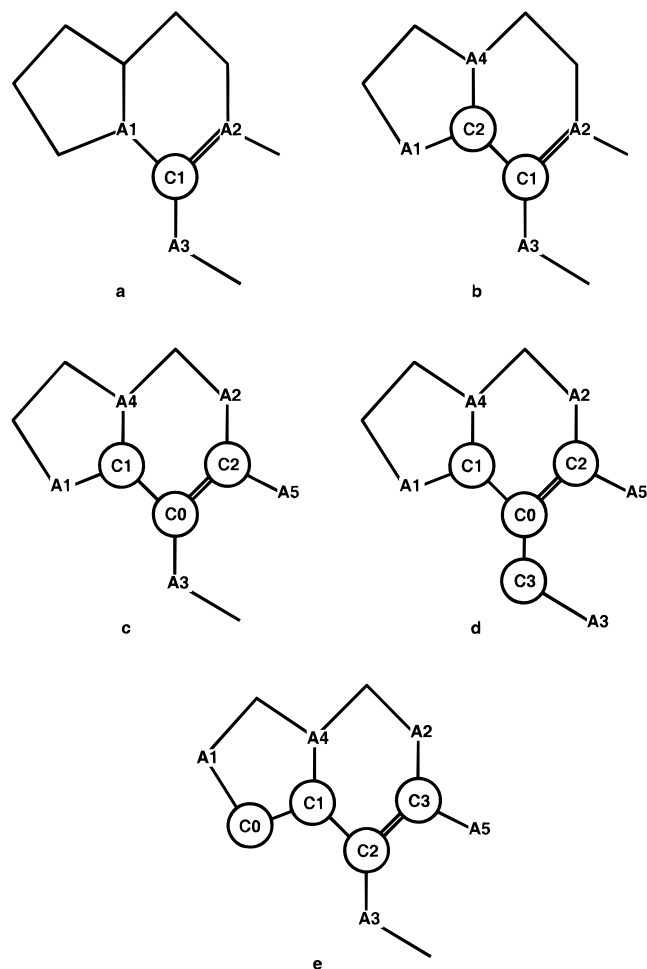


Figure 1. Definition of monocentered group (a), dicentered group (b), tricentered group (c), tetracentered group type 1 (d), and tetracentered group type 2 (e). The core atoms are circled and labeled with letter C, and the side atoms are labeled with letter A.

where 'A' and '1' represent the ASCII codes of these characters. In notation 1, a neighbor atom always follows its corresponding bond.

Expression 1 is also considered as a descriptor of core atom C1, which will be used for the notation of higher level groups.

Dicentered Group—Level 3. Any two adjacent heavy atoms, together with their nearest neighbors, define a dicentered group (Figure 1b). It is recorded by a character string with three subunits as

$$| \text{descriptor of C1} | \text{bond order} | \text{descriptor of C2} | \quad (3)$$

where the first subunit records the descriptor of core atom C1 defined as in expression 1, the second subunit records the order of bond C1–C2 with a single character, and the third subunit records the descriptor of core atom C2. The core atom with a descriptor of higher priority occupies the first subunit. The priority between two character strings is determined by comparing each pair of member characters in corresponding positions in the order from the first positions to the last. The first inequality encountered determines which of the strings has higher priority.

Tricentered Group—Level 4. Any three adjacent heavy atoms, together with their nearest neighbors, define a

tricentered group (Figure 1c). It is recorded by a character string with five subunits as

$$| \text{DC0} | \text{b1} | \text{DC1} | \text{b2} | \text{DC2} | \quad (4)$$

where subunit DC0 records the descriptor of central core atom C0; b1 records the order of bond C0–C1; DC1 records the descriptor of side core atom C1; b2 records the order of bond C0–C2; and DC2 records the descriptor of side core atom C2. The order of the side core atoms is determined according to the priorities of their descriptors and orders of the bonds linking them to the center.

Tetracentered Group—Level 5. Any four contiguous heavy atoms, together with their nearest neighbors, define a tetracentered group. There are two types of tetracentered groups. Type 1 has a central core atom and three side core atoms linked to the center (Figure 1d), while type 2 has chain-like linking as C0–C1–C2–C3 (Figure 1e). A tetracentered group of type 1 is recorded by a character string with eight subunits as

$$| 1 | \text{DC0} | \text{b1} | \text{DC1} | \text{b2} | \text{DC2} | \text{b3} | \text{DC3} | \quad (5)$$

where subunit 1 indicates that it is type 1 tetracentered group, and the following subunits record the descriptor of the central core atom and the bond orders and descriptors of the three side core atoms, respectively. While the central core atom is always located at the second subunit, the three side core atoms and their bonds should be arranged in descending order according to the priorities of their descriptors and bond orders.

A tetracentered group of type 2 is recorded with the same format as string 5, but the first subunit is filled with character '2' instead of '1' and the rest of the subunits are filled with the descriptors of the core atoms and the orders of the bonds between the core atoms. These are entered in the order C0–C1–C2–C3 if the C0 end has higher priority than the other end. The order of recording will be reversed if the C3 end has higher priority. The priorities of the ends are determined according to the descriptors of the corresponding atoms and bonds in a way that the atoms and bonds closer to the ends are compared first.

Group Type. Two *n*-centered groups with identical notations will be attributed to a same group type.

Multilevel Grouping Analysis of a Molecule. Providing the above definitions of *n*-centered groups, the local structures within a molecule can be analyzed at multiple levels. The atom types within a molecule are first assigned (level 1 analysis). Then, *n*-centered groups are searched for *n* varying from 1 to 4. The groups within each level are compared between each other, and classified into a set of unique group types. The occupancy, which corresponds to the number of appearances of a group type within a molecule, is calculated for each unique group type.

Drug Group Library. A multilevel grouping analysis is performed for each molecule in the drug library. It results in a set of group types for each molecule with corresponding occupancies. The group types from different molecules are compared to identify a set of unique group types across the library. For each unique group type, three quantities w_1 , w_2 , and f are determined, where (1) w_1 , minimum occupancy, is

the occupancy in the molecule in which a group type appears the least number of times ($w_1 = 0$ if one molecule cannot be assigned to a particular group type); (2) w_2 , maximum occupancy, is the occupancy in the molecule in which a group type appears the most number of times; and (3) f is the fraction of molecules that contain a particular group type. The unique group types, as well as the associated values of w_1 , w_2 , and f , are stored into the Drug Group Library (DGL), a database which will be used in the calculations of drug compatibility of a compound.

Drug Compatibility of a Compound. A multilevel grouping analysis is performed on a compound. Each of the identified group types is compared with each of the types in the DGL. A group type of a compound is compatible with DGL if its string notation matches one of those in the DGL and its occupancy falls between the corresponding w_1 and w_2 in the DGL. If all the group types at level n are compatible with the DGL, level n is called "a compatible level". The drug compatibility, or MLCC value, of a compound is evaluated as follows,

$$\text{MLCC} = \text{MAX}\{0, n_1, n_2, n_3, \dots, n_N\} \quad (N < 6) \quad (6)$$

where MAX is a function that returns the maximum of its argument values, and the arguments include the constant 0 and the values of all the compatible levels. Thus, the MLCC value of a compound corresponds to the highest level at which the compound is found compatible with DGL, or zero if no compatibility is found at any level. In other words, an MLCC value of 0, 1, 2, 3, 4, or 5 indicates total incompatibility, compatibility only in atom type, compatibility up to mono-, di-, tri-, or teracentered group, respectively.

Drug Library. The drug library used for calculations contains 6522 compounds from the CMC (98.1) database⁹ and 5182 from the MDDR (97.2) database,¹⁰ totaling 11704 compounds. All the compounds in CMC are used except for those labeled "antineoplastic", "anticancer", "radiopaque", "contrast agent", "solvent", "surfactant", "sunscreen", "ultra-violet screen", "emetics", "preservatives", "aerosol propellant", "chelator", or "buffers". All the compounds in MDDR are used except for those labeled "biological testing", "antineoplastic", or "anticancer". Anticancer drugs are excluded from normal drugs because they are often highly cytotoxic and are likely to react with protein targets. The compounds under "biological testing" may not be drugs. However, those passed through the biological testing phase have certain proved quality, so they were added to the drug library.

Test Compounds. The top selling drugs were determined based on the total worldwide sales in 1997 in U.S. dollar values as published by PharmaBusiness.¹¹ The anticancer drugs and protein drugs were excluded. The compounds under biological testing were those in the MDDR database which are indicated as "biological testing" in the ACTIVITY field (excluding anticancer compounds). The test set for anticancer drugs was those in CMC which are indicated as "antineoplastic". The set used to represent non-drug-like molecules include alkylating agents, acylating agents, unstable agents, carcinogens, and compounds with toxicophores such as nitrosamine, nitrosourea, hydrazine, thiourea, and phosphamine.

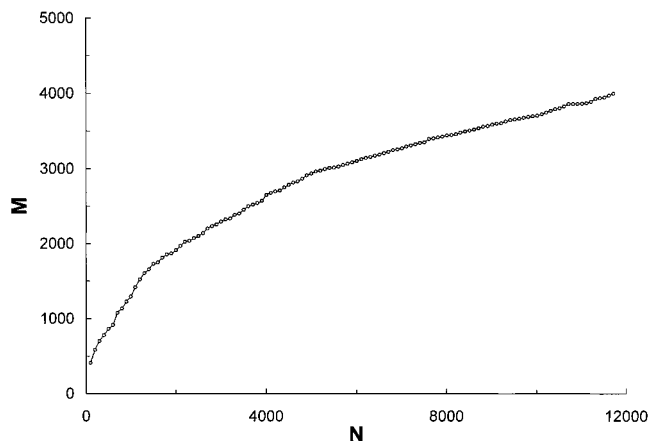


Figure 2. Plot of the number M of unique group types versus the number N of molecules collected in the accumulation of the drug group library (DGL) for dicerated groups.

3. Results

We first applied the multilevel grouping analysis to 11704 compounds in the drug library. This resulted in a set of unique group types that is saved in the DGL. This library was subsequently used to analyze the drug compatibility of the test compounds. The following sections concern the generation of DGL and the applications to the top selling drugs, compounds under biological testing, anticancer drugs, and known problematic molecules, as well as a comparison of the results from the different sets of compounds.

Completeness and Characteristics of the DGL. For each of the grouping levels, the number M of group types in the DGL is plotted as a function of the number N of molecules collected from the drug library. The plot for the dicerated group is given in Figure 2 as an example. One can see that the slope of the curve is much deeper at the beginning than at the end. The slowing down of the M variation implies that fewer new group types per molecule are identified when more and more molecules are collected. In the later phase, when a molecule is added, many of the group types of the molecule are identical to the existing types in the DGL, so that fewer unique group types are added into the DGL. Ideally, a zero growth of M would be expected at the end of collection if no new group type can be found.

While the exact zero growth was not observed, an assessment of the completeness, σ , of the DGL can be obtained based on the initial and final slopes, k_i and k_f , of the curve, using the following equation

$$\sigma = 1 - k_f/k_i \quad (7)$$

σ is equal to 1 if the zero growth is reached, or 0 if no slowing down is observed. Let us consider the ensemble of the potential group types as a closed space, and that at the final stage of collection of molecules, a fraction x of the space is populated. The probability of hitting an unpopulated spot, or new group type, in the space will be proportional to $(1-x)$. Let us also consider that a molecule generates n random group types on average. The n random group types will contribute, at the final stage, only k_f new group types, which is equal to $(1-x)n$ on a statistical basis. The n can be approximated by k_i since all group types are new at the

Table 2. Sizes and Completeness of the Drug Group Libraries

grouping level	N^a	k_i^b	k_f^c	σ^d
atom type	40			
monocentered	949	0.230	0.059	0.74
dicentered	3995	1.455	0.156	0.89
tricentered	18748	5.98	0.731	0.88
tetracentered	69646	16.5	3.25	0.80

^a Total number of atom types or group types. ^b Initial slope, dimensionless. Calculated by least squares fitting to the first 300 molecules in a collection. ^c Final slope, dimensionless. Calculated by least squares fitting to the last 1700 molecules in a collection. ^d Completeness, dimensionless.

beginning and no statistical preferences are assumed between the initial samples and the final ones. In this case, x is actually the σ defined in eq 7. Therefore, σ can be interpreted as the fraction of the group type space occupied at the end of the collection.

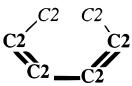
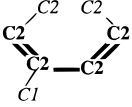
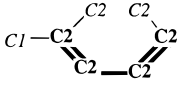
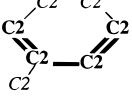
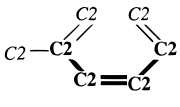
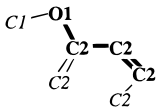
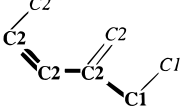
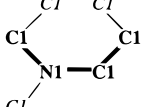
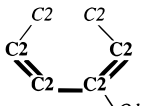
The values of k_i , k_f , and σ were calculated based on the plot of M versus N , for mono-, di-, tri-, and tetracentered groups, respectively. These are given in Table 2, together with the total number of group types for each level of grouping and for atom type. This indicates that the completeness of the current DGL is 74%, 89%, 88%, or 80% of the potential group type space, respectively, for mono- to tetracentered group.

The most common atom types in the obtained DGL are C1, C2, N1, O1, O2, N2, G1, S, F, Na, and P with frequency going in descending order. The sp^2 carbon is equally as populated as the non- sp^2 carbon. The sp^2 oxygen is slightly less populated than the non- sp^2 oxygen. The sp^2 nitrogen is significantly less populated than the non- sp^2 nitrogen. The halogen atoms G1 (Cl, Br, or I) and F are frequently found in drugs.

Table 3 gives some of the most populated tetracentered group types with the corresponding w_1 , w_2 , number, and percentage of molecules. There are usually 6 to 7 heavy atoms involved in defining a tetracentered group, sufficient to almost cover an entire phenyl ring. A tetracentered group can cover up to 14 heavy atoms in highly substituted structures. The most populated types can match groups such as phenyl and conjugated polyring, as well as alkyl, alcoholic, or amine substitutions of these groups. The structures of alkyl trisubstituted amines and nonbranched alkyl chains are highly populated. However, a compound bearing the most populated drug motifs may not necessarily be a drug. A regular MLCC calculation is needed to assess drug feasibility of a compound.

Application to the Top Selling Drugs. To answer the question as to whether the "good" drugs have high compatibility to DGL, we performed the following computer experiments. In one experiment, five top selling drugs were taken, and their drug compatibility values (MLCC) were calculated. To do this, the testing drugs were first removed from the drug library. The DGL was regenerated in the absence of the contributions of these drugs. The resultant DGL was used for calculation of MLCC of the testing drugs. After one experiment, the next five top selling drugs were taken and the same calculations were performed. Such an

Table 3. Most Populated Tetracentered Group Types in the Drug Group Library

No.	structure	w_1^a	w_2^b	n^c	f^d
1		0	7	3767	0.322
2		0	6	3054	0.261
3		0	5	2137	0.183
4		0	6	1693	0.145
5		0	8	1639	0.140
6		0	4	1336	0.114
7		0	4	1249	0.107
8		0	12	1248	0.107
9		0	4	1246	0.106

^a Minimum occupancy. ^b Maximum occupancy. ^c Number of molecules. ^d Fraction of molecules.

experiment was repeated 16 times so that 79 top selling drugs were calculated (the last experiment had only four drugs). The resulting MLCC values as well as the annual sales of each tested drug are listed in Table 4. The percentage of molecules that show compatibility at various levels are given in the first row of Table 5. It indicates that 60 out of 79 molecules are fully drug-compatible, 70 are compatible at level 4, and 100% of the molecules are compatible at the atom type level. While a majority of the molecules are drug-compatible, there are a significant number of incompatible molecules. The percentages of the incompatible molecules coincide with the percentages of the unoccupied space of group type by the DGL: 24%, 11%, or 8% versus 20%, 12%, or 11% for tetra-, tri-, or dicentered group, respectively. It

Table 4. Calculation of the MLCC of the Top Selling Drugs

no.	name	sales ^a	MLCC	no.	name	sales ^a	MLCC
1	simvastatin	3,575.0	5	41	terbinafine	628.4	5
2	omeprazole	2,815.8	5	42	doxazosin mesylate	626.0	4
3	fluoxetine	2,559.0	5	43	terazosin	620.0	5
4	enalapril	2,510.0	5	44	ondansetron	619.9	5
5	ranitidine	2,255.0	5	45	indinavir	582.0	5
6	amlodipine besylate	2,217.0	5	46	metformin	579.0	4
7	loratadine	1,726.0	5	47	propofol	568.2	5
8	amoxicillin	1,517.0	5	48	atenolol	551.9	5
9	sertraline	1,507.0	5	49	beclomethasone	542.8	5
10	paroxetine	1,474.0	5	50	itraconazole	537.0	3
11	ciprofloxacin	1,441.1	5	51	alendronate	532.0	5
12	pravastatin	1,437.0	5	52	imipenem	530.0	5
13	clarithromycin	1,300.0	5	53	cilastatin	530.0	5
14	cyclosporine	1,254.0	1	54	nizatidine	526.5	5
15	famotidine	1,180.0	1	55	divalproex	520.0	5
16	diclofenac	1,105.8	5	56	fluticasone	516.6	1
17	nifedipine	1,101.0	5	57	warfarin	500.0	5
18	lovastatin	1,100.0	5	58	nabumetone	489.0	5
19	sumatriptan	1,085.7	5	59	zidovudine	470.7	5
20	cisapride	1,045.0	5	60	benazepril	456.1	5
21	lisinopril	1,035.0	5	61	isotretinoin	451.3	5
22	ceftriaxone	1,011.4	2	62	buspirone	443.0	5
23	acyclovir	951.2	5	63	cefaclor	442.2	2
24	fluconazole propionate	881.0	2	64	midazolam	431.3	3
25	atorvastatin	865.0	4	65	ceftazidime	426.1	5
26	risperidone	848.0	5	66	fluvastatin	425.8	4
27	diltiazem	825.7	5	67	troglitazone	420.0	4
28	azithromycin	821.0	3	68	carbamazepine	414.4	4
29	captopril	795.0	5	69	metoprolol	413.6	5
30	olanzapine	730.0	5	70	clozapine	408.6	5
31	lanzaprazole	730.0	5	71	venlafaxine	403.2	4
32	ipratropium	691.7	5	72	finasteride	400.0	5
33	losartan	681.0	5	73	pentoxifylline	399.4	5
34	lamivudine	677.3	5	74	stavudine	398.0	4
35	salmeterol xinafoate	665.8	5	75	zolpidem	396.0	4
36	norethindrone	658.0	5	76	quinapril	378.0	5
37	mestranol	658.0	5	77	teprenone	367.7	4
38	cefuroxime axetil	649.4	5	78	granisetron	366.0	5
39	budesonide	643.9	5	79	ketoconazole	364.0	5
40	albuterol	641.2	5				

^a Worldwide annual sales in 1997, in million dollars.¹¹

Table 5. Comparison of the Results from Four Sets of Compounds

category ^a	N ^b	MLCC ≥ 1 (%) ^c	MLCC ≥ 2 (%) ^c	MLCC ≥ 3 (%) ^c	MLCC ≥ 4 (%) ^c	MLCC = 5 (%) ^c
drugs	79	100	96	92	89	76
bio testing	68017	99.8	97.9	88.2	59.0	27.4
anticancer	461	96.1	87.2	62.9	38.0	19.1
problematic	57	100	72	39	7.0	0.0

^a Four sets of compounds: top selling drugs, compounds under biological testing, anticancer drugs, and known problematic compounds.

^b Number of compounds in each test set. ^c Percentage of the compounds with MLCC values at or above a given value.

is likely that the incompatibility found in this case is mainly due to the incompleteness of the DGL (see Discussion section for details).

For understanding the structural features of drug compatibility, the top 10 drugs, which are also fully drug-compatible, are shown in Figure 3. The compatible drugs are often exclusively composed of building blocks that exist in natural biomolecules. These building blocks include lactone, lactam, furan, indole, imidazole, amide, alkyl-substituted amines, alkyl thioether, phenol, and guanidyl groups. In addition, halogenated aromatics are often observed in good drugs.

Application to the Compounds under Biological Testing. To examine how many of the development compounds are drug-compatible, we applied the analysis on 68017

compounds extracted from the MDDR database. These compounds were claimed to be under biological testing by various organizations. The full drug library was used in the derivation of the DGL. The MLCC values were calculated for all the compounds. The percentage of compounds that show drug compatibility are given in the second row of Table 5, as a function of the level of grouping. It indicates that 27.4% of compounds are fully drug-compatible, 59.0% are compatible at the levels up to tricentered group, and the majority of the compounds become compatible at lower levels. The chemical structures of 10 examples of drug-compatible compounds in this set are given in Figure 4. For comparison, 15 examples of incompatible compounds are given in Figure 5. It is very interesting to note that, as for

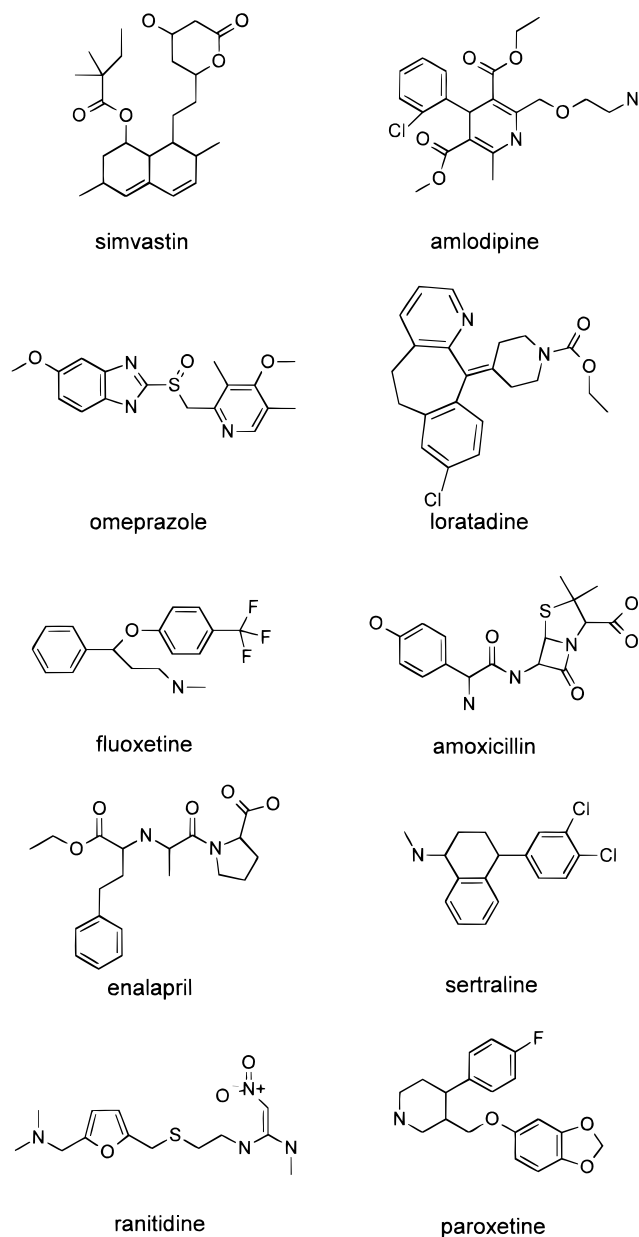


Figure 3. Chemical structures of the top 10 drug molecules.

the known drugs, the natural building blocks are also found in drug-compatible compounds: imidazole and amide in compounds **1** and **9**, steroid ring in compound **2**, indole and urea in compound **5**, alkyl-substituted amine in compound **6**, lactam and amide in compound **8**. In addition, the following groups are also observed in the drug-compatible compounds: fluorophenyl, piperidine, 1,4-dioxane, alkyl esters, piperazine, nitrophenyl, methoxyphenyl, chlorophenyl, chloropyridine, and trifluoromethylphenyl. In Figure 5, the local structures responsible for the drug incompatibility of the compounds are highlighted. The incompatible compounds contain unnatural building blocks or specifically substituted natural building blocks. Some of the incompatible local structures have apparent undesirable features. For example, the constructs of peroxide and chloromethyl ketone are most likely unstable or reactive (compounds **2** and **7**); and those of thiourea and disubstituted hydrazine may be toxic through metabolic toxication^{4,5} (compounds **5** and **6**). However, for some of the other structures, the reasons for their incompat-

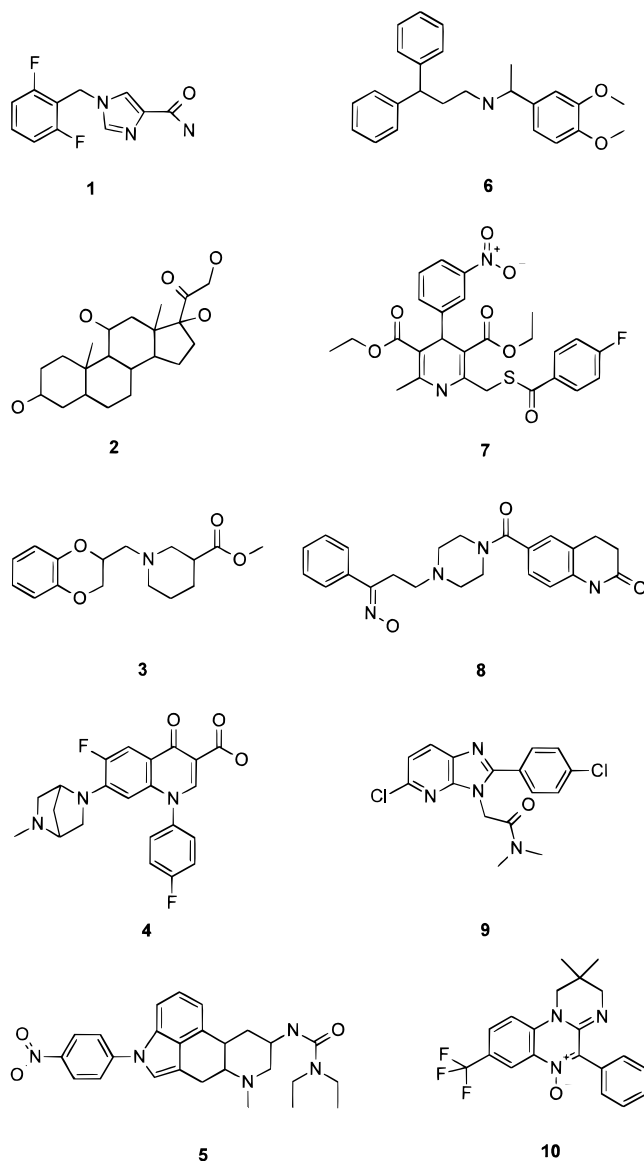


Figure 4. Chemical structures of some examples of drug-compatible molecules among those under the phases of biological testing.

ibility are unknown. For example, the incompatible structure in compound **14** indicates a β -lactam fully substituted with C2 atoms and with sp^2 hybridization at the α -position. It would be interesting to experimentally verify if and why this structure is unsuitable for drugs.

Application to Anticancer Drugs. Most anticancer drugs are cytotoxic. This type of compound should be more frequently drug-incompatible in comparison to normal drugs or drug candidates. To see if the method can successfully demonstrate this point, we applied the analysis on the 461 anticancer drugs extracted from the Comprehensive Medicinal Chemistry (CMC) database. The percentage of compounds that show drug compatibility are given in the third row of Table 5, as a function of the grouping level. It indicates that 19.1% of compounds are drug-compatible at level 5, 38.0% are compatible at level 4, and the percentage increases accordingly when the grouping level decreases. The fraction of drug-compatible individuals in anticancer drugs (19.1% at level 5) is small relative to that in the top selling drugs (76%) or in the biological testing compounds (27.4%).

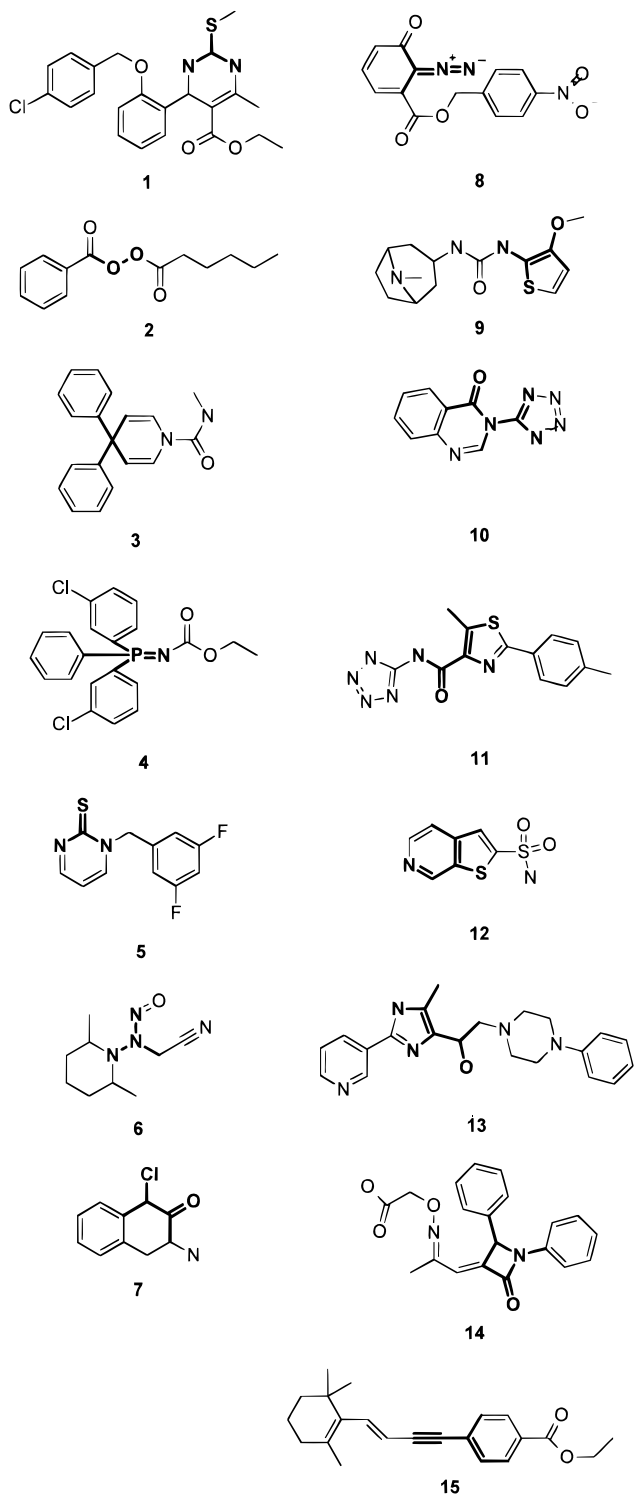


Figure 5. Chemical structures of some examples of drug-incompatible molecules among those under the phases of biological testing. The incompatible local structures are highlighted.

Our method outlined in this paper was successful in demonstrating the absolute and relative high population of toxic compounds in anticancer drugs.

Application to Known Problematic Compounds. To examine if the method can detect the commonly known toxic or reactive compounds, a set of existing compounds with known problematic local structures was collected from the Available Chemical Directory (ACD)¹² and from the books of Jakoby⁴ and Powis and Hacker.⁵ The structures of these

compounds are enumerated in Table 6 to demonstrate their non-drug-like features. The MLCC value was calculated for each of the compounds. It is given alongside each corresponding structure in Table 6. The percentage of compounds that showed drug compatibility is given in the last row of Table 5 as a function of the level of grouping. It indicates that none of the 57 compounds are drug-compatible at level 5. Only 7% are compatible at level 4. The percentage increases with decrease of the grouping level. The calculations demonstrate that the method was able to attribute the known non-drug-like molecules as non-drugs.

Comparison of Different Test Sets. In summarizing the results presented in Table 5, a general trend is observed in terms of the fraction of compounds showing drug compatibility: (top selling drugs) > (biological testing compounds) > (anticancer drugs) > (known problematic compounds). This is true for almost all the *n*-centered groupings. The majority of the compounds in the top selling drugs are fully drug-compatible. About one-quarter of the biological testing compounds are fully drug-compatible. About one-fifth of the known anticancer drugs are fully drug-compatible. None of the known problematic compounds are drug-compatible.

4. Discussion

The calculation of the multilevel chemical compatibility is deterministic. An unambiguous answer is given for the compatibility of a compound with the drug library in terms of the local structures. Owing to the deterministic feature, the problem of validation of the method is isolated into the problems of the quality of the drug library itself and the biochemical significance of the local structure compatibility. The major problem of the drug library is its incompleteness: all the potential local structures useful in constructing a drug do not exist in the current drug library. New drugs with undiscovered local motifs remain possible. Thus, a detection of incompatibility between a test compound and the DGL may be due to the incompleteness of the DGL, instead of a true "defect" of the compound. However, we demonstrated that 70–90% of the group type space has been covered by the current DGL, depending on the level of grouping. Only a small fraction of group types remains undiscovered. Therefore, incompatibility of a compound detected by this method probably reflects a true defect of the compound. In addition, a compound showing full compatibility to the current DGL is certainly a drug-compatible compound in a deterministic sense, providing the drugs in the library are true drugs.

How probable is it that a drug-compatible compound is actually a drug? And how probable is it that a drug-incompatible compound is not a drug? To explore these questions, we conducted a number of tests. The tests were designed to examine if there is any correlation between the drug eligibility of compounds and the local structure compatibility. The MLCC calculations were performed on four sets of compounds: the top selling drugs, biological testing compounds, anticancer drugs, and known problematic compounds. The top selling drugs are considered "good drugs". The biological testing compounds are the compounds in the phases preceding the clinical trials according to the publications and patents. Many of them may not be drugs. The

Table 6. Numbering, Chemical Structures, and Multilevel Chemical Compatibility (MLCC) Values of the Known Problematic Molecules

no.	mol structure	MLCC	no.	mol structure	MLCC	no.	mol structure	MLCC
1		4	20		1	39		4
2		2	21		1	40		2
3		3	22		1	41		1
4		4	23		1	42		1
5		2	24		3	43		2
6		2	25		3	44		2
7		3	26		2	45		2
8		3	27		1	46		2
9		3	28		2	47		2
10		3	29		2	48		3
11		4	30		3	49		3
12		1	31		1	50		3
13		2	32		3	51		2
14		2	33		3	52		3
15		2	34		2	53		2
16		2	35		3	54		1
17		1	36		3	55		1
18		1	37		3	56		1
19		1	38		3	57		1

anticancer drugs should be largely non-drug-like due to their cytotoxicity. The known problematic compounds are those

that are clearly not suitable for drugs (see Methods for details). The calculations on these four sets of compounds

indicate that, in terms of the fraction of compounds showing drug compatibility, (top selling drugs) > (biological testing compounds) > (anticancer drugs) > (known problematic compounds). More drug-compatible compounds are found in the sets containing more drug-like molecules. Therefore, the method is able to capture the trend in feasibility of compounds as drug candidates for the tested sets.

The calculations indicated that the majority of the top selling drugs are fully drug-compatible in terms of the local structures. About one-quarter of the biological testing compounds are fully drug-compatible. About one-fifth of the anticancer drugs are fully drug-compatible. None of the known problematic compounds are drug-compatible. These results are consistent with the status of the compounds in each set. However, a significant fraction of top selling drugs (24%) are not fully drug-compatible. This may be due to the incompleteness of the drug library.

The compatibility of a compound to the DGL depends on the level of grouping. The higher the level of grouping, the more the nonlocal chemical features are captured, and the more selective the method is. However, too high a level of grouping may multiply the diversity of group types so that the completeness of the library becomes a serious problem. According to the above tests, most known non-drug-like molecules are removed at the tricentered group level and all the non-drug-like molecules are removed at the tetracentred group level. On the other hand, a substantial number of the top selling drugs start to be removed at the tetracentred group level. As a compromise, we suggest that the tri- or tetracentred group level is chosen as the threshold for determining the drug compatibility of a compound.

The current algorithm requires that every group type within a molecule has to be identical to some part of an existing drug, if a molecule is to be ranked as a drug. This may cause an over-discriminative prediction: drugs with new, but acceptable, group types are ranked as non-drugs. A remedy of this potential problem would be the introduction of certain similarity measures in determining compatibility between a target group type and a drug group type. A pair of different group types could be ranked "compatible" if they were similar in nature. Thus, the molecules with new, but similar, group types would also be ranked as drugs.

Examination of the chemical structures allows one to clarify which types of structures are drug-compatible and which types are not. Many incompatible local structures found correspond to the groups that are likely to be either toxic or reactive. This confirms the capacity of our method for identifying non-drug-like compounds. A general statement can be made concerning the structural features. The drug-compatible compounds are often composed of building blocks that exist in natural biomolecules, such as lactone, lactam, and amino acid side chain analogues. These building blocks, however, are assembled somewhat differently in the drugs than in the natural biomolecules. In contrast, the drug-incompatible compounds often contain unnatural building

blocks, or natural building blocks with specific substitutions. The specific substitutions or modifications on a natural building block may lead to non-drug-like compounds in certain situations.

Direct application of the MLCC method to virtual combinatorial chemistry libraries allows to "cherry-pick" a subset of individual compounds with drug-like features. Straightforward modifications can be made on the current algorithm to select compounds in such a way that the selected candidates are accessible through combinatorial synthesis.

Summarizing the above analysis, the multilevel chemical compatibility approach is able to determine the compatibility between a compound and a drug library in terms of the local structures. The applications on the four sets of chemicals demonstrate that the compatibility in local structure implies drug-like character to a large extent. It is appropriate to use MLCC as an expression of the drug-like character of a compound. The method is fast enough to be used for prescreening of virtual combinatorial chemistry libraries and databases of compounds.

Acknowledgment. We thank Drs. Darryl Rideout, Christina Niemeyer, Cindy Fisher, and Shankari Mylvaganam for reading of the manuscript.

References and Notes

- (1) Fecik, R. A.; Frank, K. E.; Gentry, E. J.; Menon, S. R.; Mitscher, L. A.; Teliikepalli, H. The search for orally active medications through combinatorial chemistry. *Res. Rev.* **1998**, *18*, 149–185.
- (2) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and non-drugs. *J. Med. Chem.* **1998**, *41*, 3325–3329.
- (3) Ajay, W.; Walters, P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "non-drug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314–3324.
- (4) Jakoby, W. B. *Enzymatic Basis of Detoxication. Volume I*; Academic Press Inc.: Orlando, 1980.
- (5) Powis, G.; Hacker, M. P. *The Toxicity of Anticancer Drugs*; Pergamon Press: New York, 1991.
- (6) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure–activity relationships. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.
- (7) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (8) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
- (9) Comprehensive Medicinal Chemistry Release 98.1 is available from MDL Information Systems Inc., San Leandro, CA 94577. An electronic database of *Comprehensive Medicinal Chemistry* published by Pergamon Press in March 1998 contains drugs already in the market.
- (10) MDL Drug Data Report is available from MDL Information Systems Inc., San Leandro, CA 94577. An electronic database version of the Prous Science Publishers journal *Drug Data Report*, extracted from issues starting mid-1988, contains biologically active compounds with information on their development and clinical trials.
- (11) *Annual Report*, PharmaBusiness No 22, 1998-07-00, page 38+.
- (12) Available Chemical Directory, version 97.1, is an electronic database of commercial available compounds, available from MDL Information Systems Inc., San Leandro, CA 94577.