Finding Multiactivity Substructures by Mining Databases of Drug-Like Compounds

Robert P. Sheridan*

RY50S-100, Merck Research Laboratories, Rahway, New Jersey 07065

Received January 31, 2003

We have developed a method, given a database of molecules and associated activities, to identify molecular substructures that are associated with many different biological activities. These may be therapeutic areas (e.g. antihypertensive) and/or mechanism-based activities (e.g. renin inhibitor). This information helps us avoid chemical classes that are likely to have unanticipated side effects and also can suggest combinatorial libraries that might have activity on a variety of receptor targets. The method was applied to the USPDI and MDDR databases. There are clearly substructures in each database that occur in many compounds and span a variety of therapeutic categories. Some of these are expected, but some are not.

INTRODUCTION

One very desirable property of a drug is specificity. We would prefer that the drug interact with its intended receptor but not with other receptors, lest it give rise to side effects or toxicities. On the other hand, under certain circumstances, such as combinatorial library design, we may want to invent a class of molecules containing a common scaffold or "privileged structure" that could potentially interact with a variety of receptors.[1−4] Thus, it would be useful to identify chemical structures that are associated with multiple activities. The structures might be called "privileged"[1−5] or perhaps "promiscuous"[6] depending on the context. Here we will use the neutral term "multiactivity substructures". This is for two reasons. First, we avoid implying whether having many activities is desirable or not. Second, we want to include in vivo biological effects as well as in vitro measures such as binding to a particular receptor, whereas the words "privileged" and "promiscuous" are usually in the context of the latter only. In addition, "privileged structures" is sometimes applied specifically to substructures in molecules that bind to G-protein coupled receptors. We feel "multiactivity substructures" is appropriately general.

In this paper we present a method for identifying multiactivity substructures by mining databases of drug-like molecules and their associated activities. This method is applied to the USPDI[7] (United States Pharmacopeia Drug Index) and the MDDR[8] (MDL Drug Data Report). There are clear examples in either database of substructures that are associated with many different activities or therapeutic areas. Some are expected, but some are not.

METHODS

We have the following data requirements:

1. A large set of molecule identifiers and one or more activities associated with each identifier.

2. A set of connection tables, one associated with each molecular identifier.

Our mining procedure is the following:

1. Preprocess the connection tables.

2. Preprocess the activity records.

3. Identify pairs of molecules that have similar structures and dissimilar activities.

4. For every pair of molecules from step 3, find the highest-scoring common substructure (HSCS). Keep only those HSCSs that are statistically significant.

5. For every molecule, generate a "consensus substructure" from it and its HSCS-neighbors (see below). Note the unique activity records for the molecule and its neighbors.

6. Sort the consensus substructures.

7. Remove redundant consensus substructures.

8. Inspect the activities associated with consensus substructures.

Preprocessing the Connection Tables. Salts are removed by keeping the largest fragment for every connection table. Since we are interested in drug-like molecules, we remove any molecules whose largest fragment contains less than 10 or more than 50 non-hydrogen atoms. For the purposes of finding common substructures, we need to assign a "type" to each atom. Here the type is a string consisting of the element concatenated with its hybridization. For instance a methyl carbon would be "C_sp3".

Preprocessing the Activity Records. The specific nature of activity records depends on the database, but generally the following are true: There may be more than one activity associated with a molecule. Activity records may indicate a general therapeutic area (e.g. "antihypertensive"), a specific receptor (e.g. "angiotensin AT1 blocker"), or be merely descriptive of the chemical structure (e.g. "biphenyl-containing compound"). A general issue with activity databases is that each molecule has been tested for only a few activities, so there are probably many molecules that falsely lack an activity record that they might have had if only they were tested in that area. Also, molecules with the same therapeutic area might work by completely different mechanisms, and it is debatable whether they should be considered as having the same activity.

There are also consistency issues with the activity records. Some activities may be nearly synonymous (e.g. "antihypertensive" vs "blood-pressure-lowering agent") and some

* Corresponding author phone: (732)594-3859; fax: (732)594-4224; e-mail: sheridan@merck.com.

10.1021/ci030004y CCC: $25.00
Published on Web 04/09/2003

are clearly subsets of others (e.g. "angiotensin AT1 blocker" vs "angiotensin blocker"). The curators of drug-like databases have made some attempt to standardize the activities, but casual inspection shows that there are large gaps in internal consistency. For instance, not all "angiotensin AT1 blockers" are also "angiotensin blockers". We feel that the problem of imposing more consistency among several hundred unique activity records or finding partial equivalences among the activities is not solved at present, so here we treat each activity record as its own unique string and leave the decision about whether two activities are truly different to the inspection phase. The only processing is to replace embedded blanks, punctuation marks, etc. in the activity records with "_". This prevents such characters from confusing our various string manipulating tools. Also, all letters are changed to uppercase.

Identify Pairs of Molecules That Have Similar Structures and Dissimilar Activities. Given an ideal database, where each molecule had a record for all of its activities, one could look for multiactivity substructures by finding individual molecules with the most records. However, in real databases each molecule is tested in only a few areas and at best has only a handful of records. One approach to building up a large set of activities is to look for pairs of similar molecules that are listed as having different activities.

We look at all pairs of molecules in the database and keep those pairs where

1. The topological similarity of the molecules using the AP ("atom pair") descriptor and Dice similarity definition is ≥ 0.7. (Details are in ref 9.) Molecules similar at 0.7 would be regarded as clear analogues by most chemists.

2. The molecules have no activity records in common.

This has three desirable effects. First, the requirement that the molecules be close analogues removes from consideration common substructures that are small relative to the size of the molecules. Second, we preselect the set of compounds most likely to contain many different activities. Third, we need to calculate common substructures for only a small subset of the total pairs. This saves a great deal of time, since generating common substructures is computationally very expensive compared to calculating Dice similarity.

Find Significant HSCSs. For this step we use the maximum common substructure method in Sheridan and Miller.[10] This method, based on clique detection, can generate substructures that are disconnected. A clique-defined substructure is a set of pairs of atoms one from molecule A and one from molecule B such that the paired atoms are of the same type and the through-bond distances between the atoms in A are the same as the corresponding distances in B. The score of a common substructure is

$$\text{score} = \text{size} - p(N_{\text{frag}} - 1)$$

where size is the number of atoms in the common substructure and N_{frag} is the number of disconnected fragments in the common substructure. p is a penalty for the substructure being disconnected. We keep only the highest scoring common substructure (HSCS) for each pair of molecules.

Any arbitrary pair of molecules is likely to have something in common, and we wish to keep only those HSCS that are much larger than expected for two randomly selected molecules of the same size, i.e., those above "noise". We define a Z-score as

$$Z = \frac{(\text{score} - \text{mean})}{\text{stdev}}$$

The expected mean and standard deviation score for two randomly selected molecules is a linear function of the number of atoms in the smaller molecule (see ref 10 for details)

$$\text{mean} = M_{\text{mean}}\min(n_A,n_B) + B_{\text{mean}}$$

$$\text{stdev} = M_{\text{stdev}}\min(n_A,n_B) + B_{\text{stdev}}$$

and n_A is the number of atoms in molecule A. For the atom types used here and $p = 1$, the appropriate slopes and intercepts for the linear relationships are $M_{\text{mean}} = 0.24$, $B_{\text{mean}} = 2.29$, $M_{\text{stdev}} = 0.051$, $B_{\text{stdev}} = 0.858$.

If $Z \geq 4$, we consider the HSCS "significant". Others are discarded. This removes from consideration smaller substructures (typically <13 non-hydrogen atoms) that can be found in unrelated drug-like molecules (indoles, biphenyls, piperazines, etc.).

Generating a Consensus Substructure. Each molecule may have one or more "HSCS-neighbors" with which it shares a significant HSCS. Each atom in a molecule is given one point for each time it appears in a significant HSCS. The maximum value for a molecule is the number of HSCS-neighbors for that molecule. If the value on an atom divided by the maximum value is above a certain threshold (here 0.5), that atom is considered "conserved". A consensus substructure (equivalent to what we called a "conserved atom display" in our original publication [10]) is created by modifying the molecule: Nonconserved atoms are marked with X and bonds to them are drawn as dotted lines. Nonconserved atoms not on a bond path between two conserved atoms are deleted. An example is shown in Figure 1 where the molecule doxepin has been transformed to a consensus substructure C_doxepin that represents the atoms conserved among doxepin and its neighbors. Figure 1 also shows the 15 unique activities among doxepin and its neighbors.

Sorting the Consensus Substructures. To be interesting a consensus substructure would

1. Occur many times in the database, i.e., have many HSCS-neighbors.

2. Have many unique activities.

One simple, robust way to take both criteria into account simultaneously without applying arbitrary coefficients is to do the equivalent of a Borda count:[11] Sort the consensus substructures by decreasing number of neighbors and note the rank of each (the one with the most neighbors is rank 1, the next rank 2, etc.). Similarly for the number of unique activities. For every consensus substructure compute the mean of the two ranks. Sort the consensus substructures by increasing mean rank.

Remove Redundant Consensus Substructures. At this point we have many consensus substructures, many of which are very similar.

We can define the topological similarity of consensus substructures in an analogous way to how the similarity of molecules is defined in step 3. We construct modified atom-pair descriptors for the molecule of the form AT1-distance-

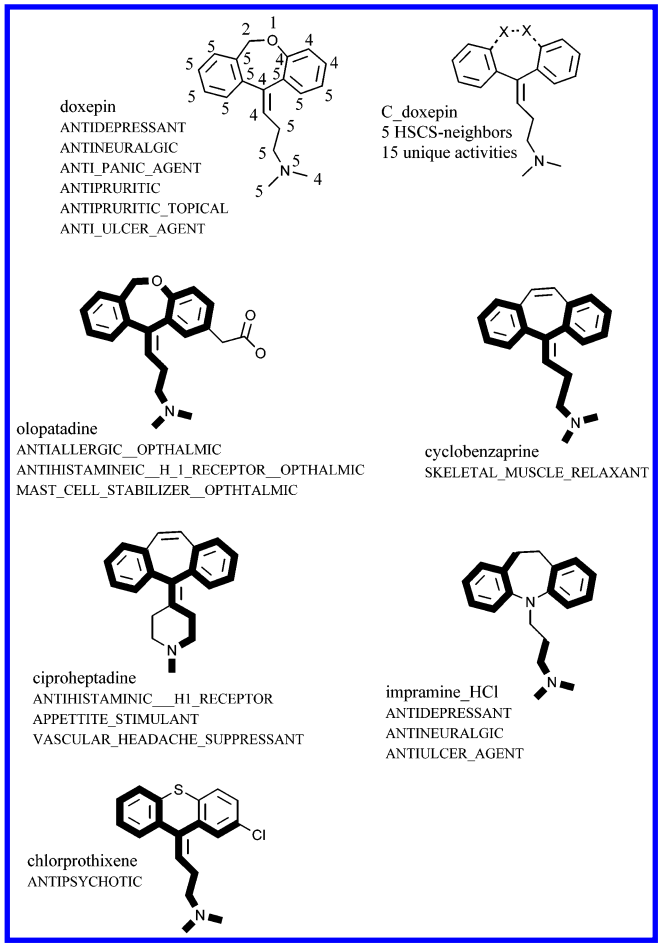


Figure 1. The HSCS-neighbors of doxepin. In each of the neighbors, the atoms that are shared with doxepin in the highest scoring common substructure (HSCS) are connected by bold bonds. (Part of the phenyl ring on the right of chlorprothixene is not matched in our clique-based method because the change in the central ring from 7 to 6 atoms changes the through-bond distances to some of the phenyl atoms.) Each atom of doxepin has been labeled with the number of neighbors with which that atom is shared. C_doxepin is a "consensus substructure" derived from doxepin such that only those atoms that are shared with the majority of its neighbors ("conserved") are kept as is. "X" marks a nonconserved atom on a path between two conserved atoms. Other nonconserved atoms are deleted (no examples in doxepin). There are 15 unique activities associated with doxepin and its five neighbors.

AT2, where AT1 and AT2 are atom types and distance is the shortest bond distance between the two atoms. The atom type of a conserved atom is inherited from the molecule (element and hybridization), and the atom type for a nonconserved atom is "X". Two consensus substructures are considered similar if their Dice similarity using the modified atom pair descriptors is ≥ 0.7.

We eliminate redundant consensus substructures in the following way: Given the Borda sorted list, keep the first consensus substructure, and eliminate those further down the list that are more similar than a given threshold. Find the next consensus substructure that has not already been eliminated. Eliminate those further down the list that are similar, etc. This leaves many fewer consensus substructures to inspect. An obvious alternative to looking at individual "best" unique consensus substructure from a set of similar ones is to agglomerate all the similar consensus substructures and look at the aggolmeration, which could contain new

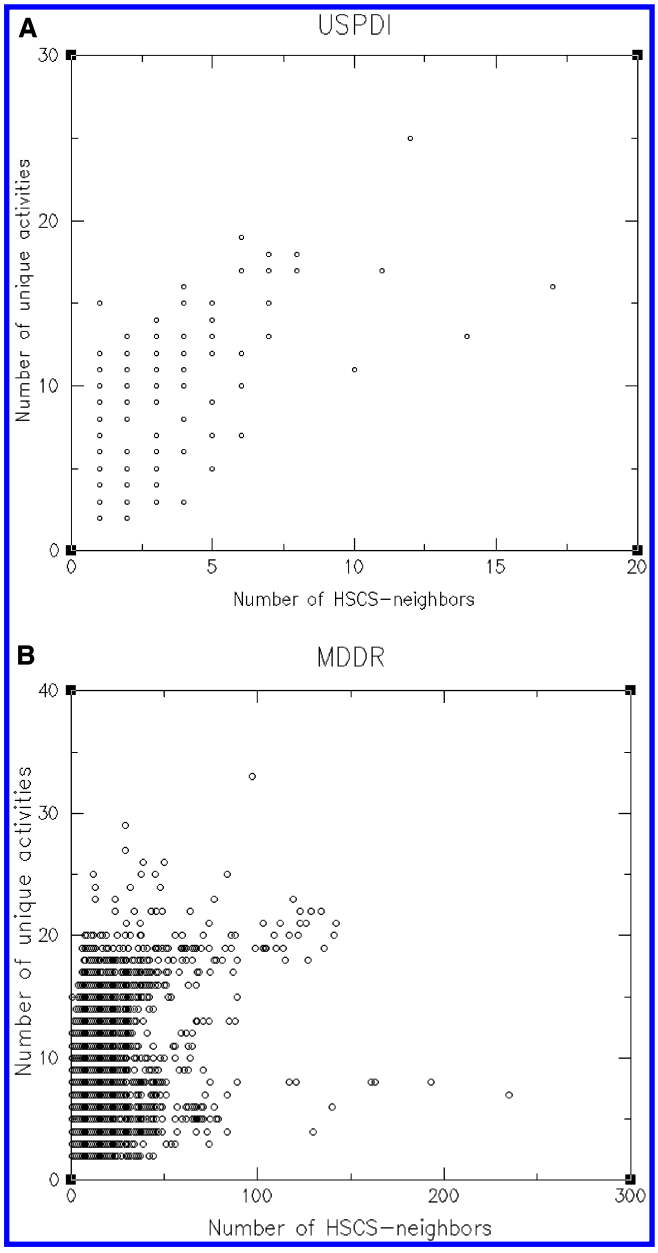
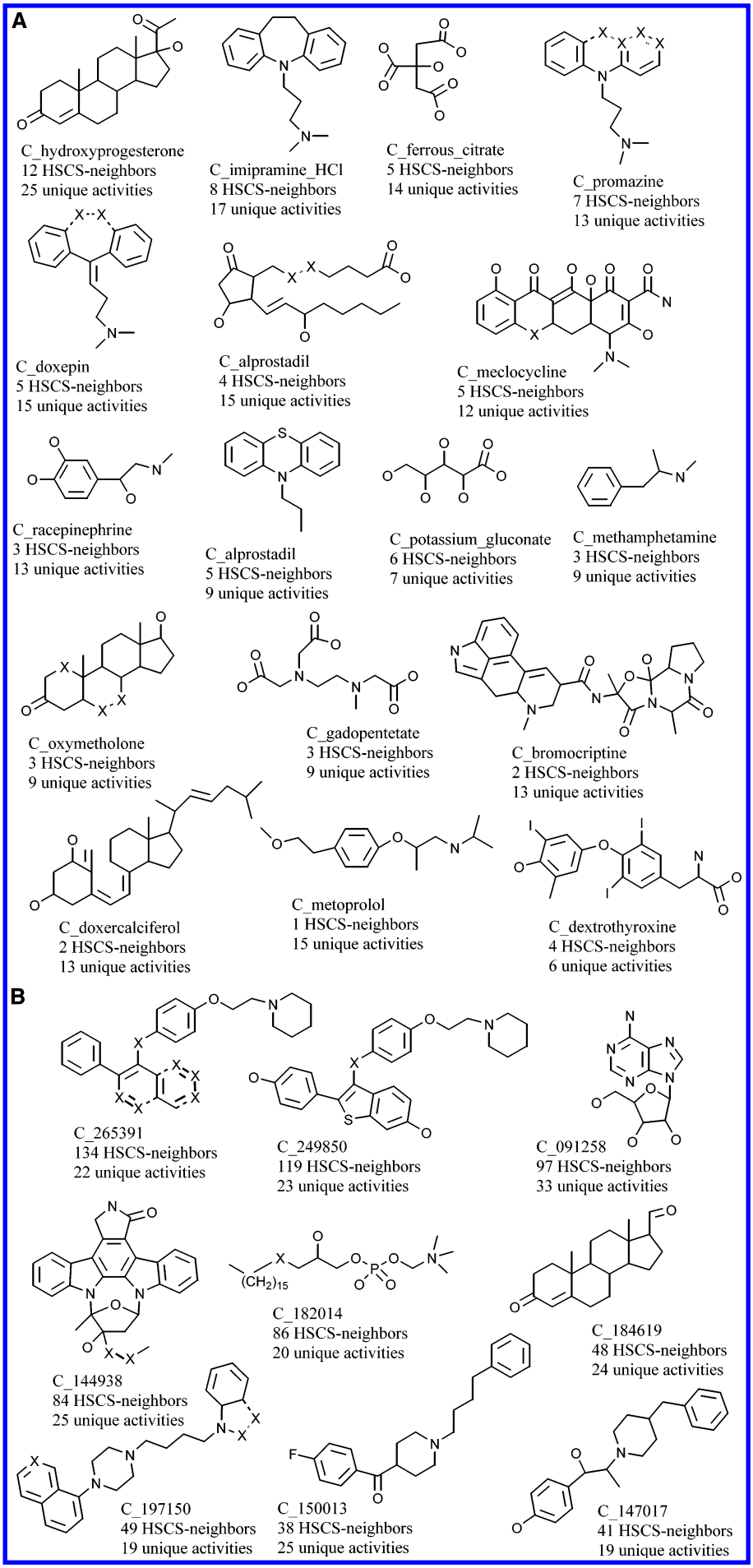


Figure 2. A plot of the number of unique activities against the number of HSCS-neighbors. A. USPDI and B. MDDR.

activities that can be missed by looking at a single consensus substructure, which in turn is already an agglomeration of several molecules. For the present, however, we will avoid this complication.

Inspection of Consensus Substructures and Their Activities. One method of inspection is to list the unique activities associated with a particular consensus substructure and the number of molecules associated with each. Another way is to produce an average linkage clustering dendrogram of the neighbors based on the Dice similarity of their descriptor records. That is, molecules would have a similarity of 1.0 if they shared all their activity records in common and 0.0 if they shared none. Interpretation of the activity records was aided by the appropriate "ACTION" records in the MDDR. Another good source is Goodman and Gilman's *The Pharmacological Basis of Therapeutics*.[12]

Databases. Here we will use the United States Pharmacopeia Drug Index (USPDI) and the MDDR (MDL Drug Data Report). USPDI[7] (June 2002 version) was licensed from


A
C_hydroxyprogesterone
12 HSCS-neighbors
25 unique activities
C_imipramine_HCl
8 HSCS-neighbors
17 unique activities
C_ferrous_citrate
5 HSCS-neighbors
14 unique activities
C_promazine
7 HSCS-neighbors
13 unique activities
C_doxepin
5 HSCS-neighbors
15 unique activities
C_alprostadil
4 HSCS-neighbors
15 unique activities
C_meclocycline
5 HSCS-neighbors
12 unique activities
C_racepinephrine
3 HSCS-neighbors
13 unique activities
C_alprostadil
5 HSCS-neighbors
9 unique activities
C_potassium_gluconate
6 HSCS-neighbors
7 unique activities
C_methamphetamine
3 HSCS-neighbors
9 unique activities
C_oxymetholone
3 HSCS-neighbors
9 unique activities
C_gadopentetate
3 HSCS-neighbors
9 unique activities
C_bromocriptine
2 HSCS-neighbors
13 unique activities
C_doxercalciferol
2 HSCS-neighbors
13 unique activities
C_metoprolol
1 HSCS-neighbors
15 unique activities
C_dextrothyroxine
4 HSCS-neighbors
6 unique activities
B
C_265391
134 HSCS-neighbors
22 unique activities
C_249850
119 HSCS-neighbors
23 unique activities
C_091258
97 HSCS-neighbors
33 unique activities
C_144938
84 HSCS-neighbors
25 unique activities
C_182014
86 HSCS-neighbors
20 unique activities
C_184619
48 HSCS-neighbors
24 unique activities
C_197150
49 HSCS-neighbors
19 unique activities
C_150013
38 HSCS-neighbors
25 unique activities
C_147017
41 HSCS-neighbors
19 unique activities

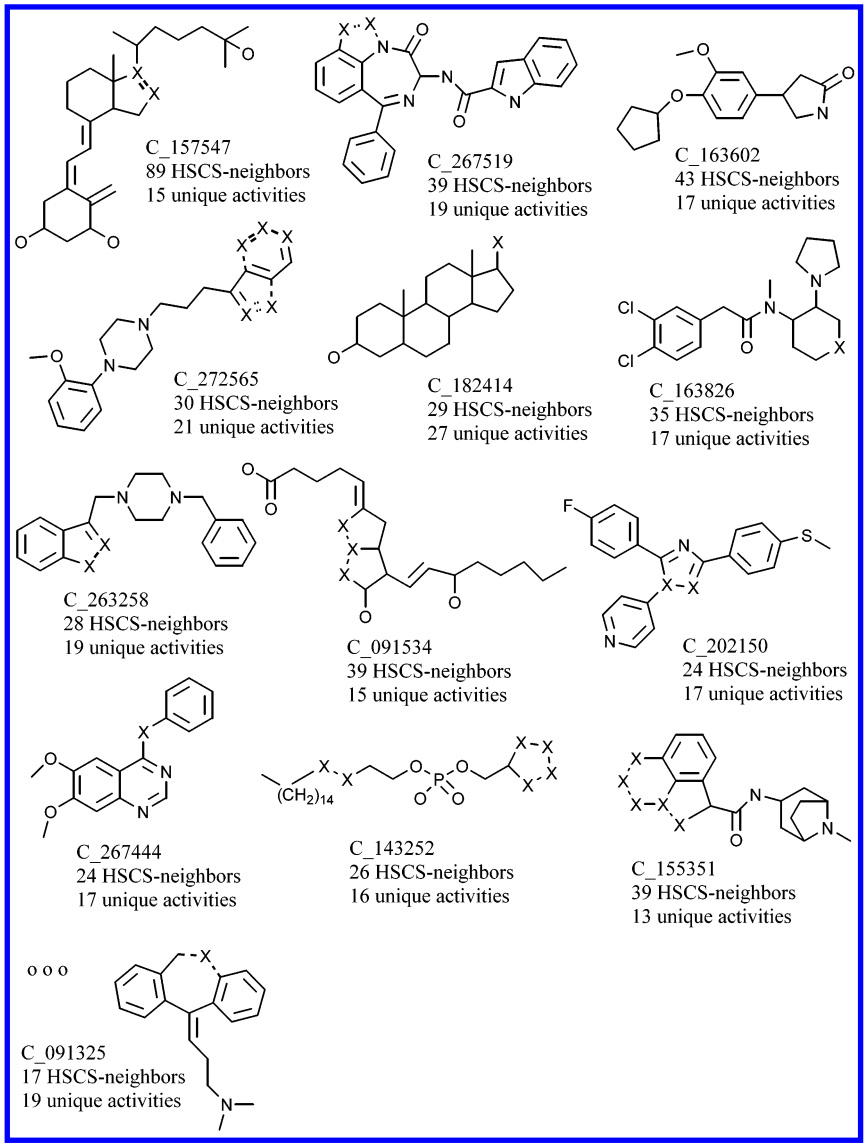


Figure 3. The best ranking consensus substructures after Borda sorting on the number of unique activities and the number of neighbors, and after eliminating similar consensus substructures. Note that these are not entire molecules but the conserved part of the named molecule relative to its HSCS-neighbors. Activities for selected consensus substructures are in Table 1. A. USPDI and B. MDDR.

Micromedex (www.micromedex.com). It contains 1167 total generic drug names with at least one associated activity. There are 591 unique activities. One complication of the USPDI is that no connection tables are provided. We needed to extract the generic names of the compounds and use our in-house program CKB[13] to assign a structure to each generic name.

MDDR[8] (version 2000.2) was licensed from Molecular Design Ltd. (www.mdli.com). It is compiled from the patent literature. It contains 119 110 six-digit compound identifiers with at least one activity and an associated connection table. There are 703 unique activities in the form of a 5-digit activity code (e.g. "71000") and a brief description (e.g. "ANTIVIRAL"). These were concatenated to produce a single activity record ("71000_ANTIVIRAL").

We feel it is useful to mine both the USPDI and MDDR. The USPDI has relatively few compounds, but since it represents a hand-curated consensus of the medical literature, each molecule almost certainly has the activity claimed for it. Almost all the activities are therapeutic areas. In contrast, the activities in the MDDR are based on patent claims, which may not be as reliable, but there are many more compounds and they are much more diverse. Also the MDDR lists mechanism-based activities related to specific receptors or enzymes.

RESULTS

The USPDI had 338 suitable pairs from step 3. There were 301 significant HSCSs among a total of 234 molecules. The equivalent numbers for MDDR are 66 783 pairs, 55 246 significant HSCSs among 22 400 molecules. The distribution of number of unique activities vs HSCS neighbors is shown in Figure 2A,B.

When the corresponding consensus substructures were Borda sorted and filtered, there were 53 and 11241 remaining in the USPDI and MDDR, respectively. The best ranked remaining consensus substructures for the USPDI and MDDR are in Figure 3 (parts A and B, respectively). Selected lists of activities are in Table 1A,B.

There are a great many interesting substructures here, and we can discuss only a few of them. It is not surprising that

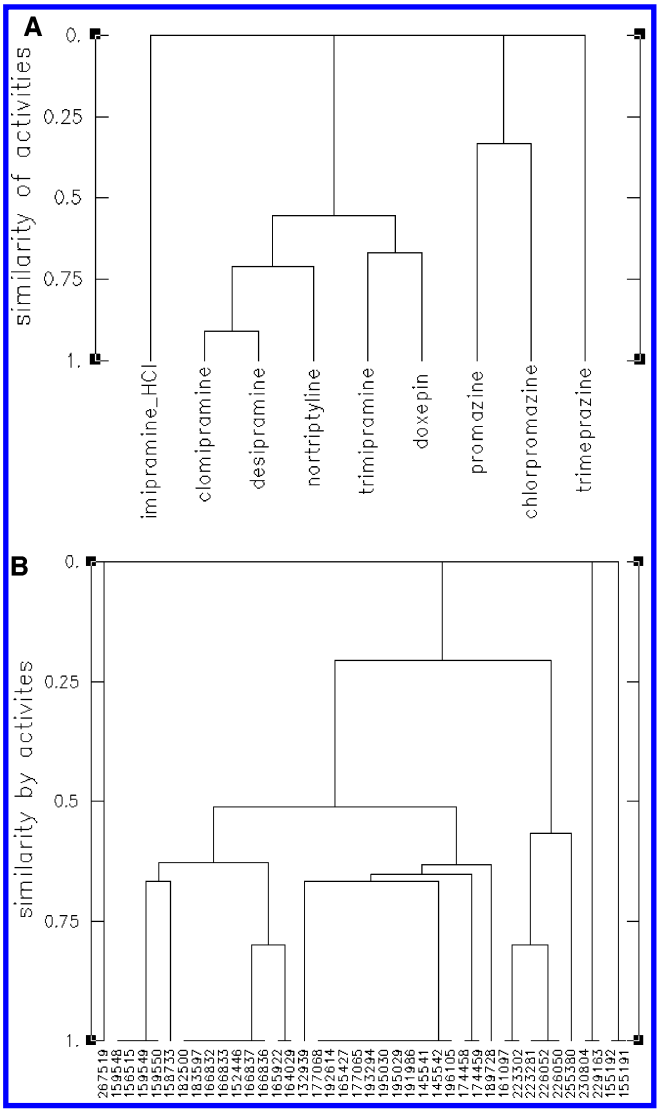


Figure 4. A dendrogram of the molecules associated with example consensus substructures. The dendrogram is based on Dice similarity of the shared activities of the compounds. For instance if compound C1 has activities A1, A2, A3, and compound C2 has activities A2, A4 their similarity would be $1/[0.5(3 + 2)] = 0.4$, i.e. one activity in common divided by the average number of activities of the two. The purpose of the dendrogram is to point out which molecules have unusual activities relative to the others. The consensus substructure is guaranteed to have a zero similarity with the others because, in the initial selection of pairs, its source molecule was chosen to have no activities in common with its neighbors. A. C_imipramine_HCl, B. C_267519.

a steroid (C_hydroxyprogesterone) is the best ranked molecule in the USPDI since steroids have a number of therapeutic effects on a number of different physiological systems.

Tricyclic antidepressant structures (C_imipramine, C_promazine, and C_doxepin) are among the best ranked. They have a number of CNS effects because they can bind to a number of CNS receptors (dopaminergic, alpha or beta adrenergic, etc.). They are also antihistamines and antipruritics (anti-itching) agents because they act at peripheral H-1 receptors. Their actions as antienuresis (anti-bed-wetting) agents are presumably due to their actions on the muscarinic or adrenergic receptors in the bladder and urethra. An activity dendrogram of imipramine_HCl is shown in Figure 4A. Imipramine is listed as an ANTIENURETIC. Clomipramine, desipramine, nortrypyline, trimipramine, and doxepin all share ANTIDEPRESSANT as an activity. Promazine and chlorpromazine share ANTIPSYCHOTIC. Trimeprazine has ANTIHISTAMINIC___H1_RECEPTOR and SEDATIVE_HYPNOTIC.

C_ferrous_citrate is an example of a molecule that has a large number of apparent activities, but on inspection almost all are related to the ability of citrate to chelate metals. It is a surprise that there are so many citrates in the USPDI.

It is not surprising that prostanoids are represented (C_alprostadil). These have various vasodilating and constricting effects.

For the MDDR, it was unexpected that the consensus substructure represented by C_265391 and C_249850 should be the best ranked. Many compounds containing a similar substructure have many activities (75000_ANTINEOPLASTIC, 50060_BONE_RESORPTION_INHIBITOR, etc.) that are probably related to their being estrogen receptor agonists or antagonists (like raloxifene). However, there are many claims for CNS-related activities (e.g. 06200_ANXIOLYTIC, 08000_ANTIDEPRESSANT). These could be due to the presence of the cation and aromatic ring. It is interesting that some of these molecules are also thrombin-inhibitors (37300_THROMBOLYTIC, 37110_THROMBIN_INHIBITOR).

The adenine nucleoside C_091258 has a number of activities related to DNA synthesis in tumors, parasites, or viruses and nucleoside metabolism (71000_ANTIVIRAL, 75000_ANTINEOPLASTIC, 72000_ANTIPROTOZOAL, etc.). There are a number of cardiovascular effects as well (28000_CARDIOTONIC, 29000_ANTIARRHYTHMIC, 31300_VASODILATOR, etc.) mediated by the adenoreceptor.

Compounds such as C_144938 probably exert antineoplastic effects by virtue of being protein kinase inhibitors, among other mechanisms. However, there are a number of CNS effects (09200_COGNITION_DISORDERS___AGENT_FOR, 111000_ANTIPARKINSONIAN, 10000_ANTICONVULSANT).

C_182014 is another surprise. There are a number of cytotoxicity claims (75000_ANTINEOPLASTIC, 71000_ANTIVIRAL) for similar alkylphospholipids, some PAF-inhibitor claims (31000_ANTIHYPERTENSIVE, 27261_PAF_ANTAGONIST), and some other enzyme inhibition claims (78348_PHOSPHOLIPASE_A2_INHIBITOR, 71522_REVERSE_TRANSCRIPTASE_INHIBITOR).

C_184619 is the first steroid in the MDDR sorted list, with a variety of activities, some of which are related to a specific receptor or enzyme (75721_AROMATASE_INHIBITOR, 78335_STEROID___5ALPHA_REDUCTASE_INHIBITOR, 78446_CYTOCHROME_P450_OXIDASE_INHIBITOR).

C_197150 has a number of CNS receptor activities: dopaminergic, cholinergic, 5-HT, anxiolytic, etc., although it may also have peripheral antihistamine activity.

Benzodiazepines have number of activities besides the central GABA/benzodiazepine receptor. C_267519 demonstrates that many activities of 2-aminobenzodiazepine derivatives are associated with the cholecystokinin (CCK) and related gastrin receptors in the gut, although there are a number of CCK-mediated CNS activities as well. A dendrogram of the activities of 267519 and its neighbors is in

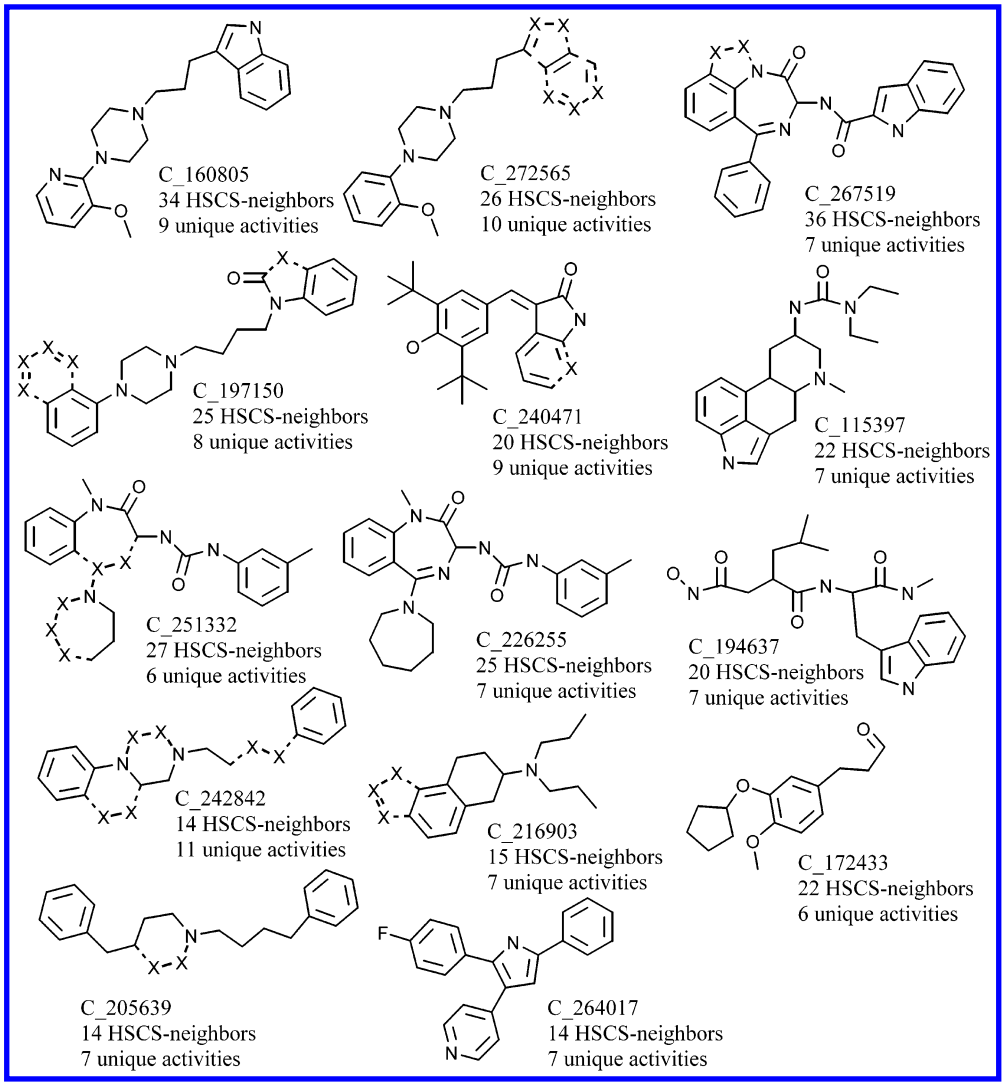


Figure 5. The best ranking consensus substructures in MDDR after Borda sorting. Here the number of neighbors and the number of unique activities takes into account only "mechanism-specific" activities. Activities for selected consensus substructures are in Table 2.

Figure 4B, which indicates that most of the compounds have related activities having to do with CCK/gastrin, but there are some outliers. 267519 itself is 27100_BRONCHIODILATOR/27200_ANTIALLERGY/ANTIASTHMATIC/78418_PHOSPHODIESTERASE_IV_INHIBITOR. 155191 and 155192 have the activity 09200_COGNITION_DISORDERS___AGENT_FOR.

Prostanoids occur with C_091534. The most common activity is 16000_ANTIGLAUCOMA, found for 19 of the 39 neighbors. The next most frequent activity is 57300_HEPATOPROTECTANT, which is found in 11 neighbors.

Tricyclic antidepressants (C_091325) appear as the 39th ranked compound in the MDDR. These show antidepressant and antihistamine-related activities, consistent with the results from the USPDI, plus miscellaneous CNS activities.

An alternate way of examining the MDDR is to consider only mechanism-based activities. The idea is that we could be more certain whether a particular consensus substructure had activities that were truly different if we knew the specific mechanism. Also, we would remove the complication with therapeutic area activities that molecules may work by different mechanisms. We selected the subset of activities that contained the words "AGONIST", "ANTAGONIST", "BLOCKER", "RECEPTOR", or "INHIBITOR". While this does not remove all the therapeutic area activities, the remaining set of 287 activities is predominantly mechanism-based. We redetermined the number of neighbors and unique activities associated for each molecule, counting only those neighbors with at least one mechanism-based activity, regenerated the consensus substructures, and resorted them. The best ranked consensus substructures are in Figure 5, and the associated activities for some of them are in Table 2.

Most of the activities are associated with G-coupled protein receptors (GPCRs). This is not so surprising because of the large amount of work characterizing GPCR subtypes. Many of the consensus substructures have activities associated with biogenic-amine binding GPCRs, for example C_160805 is associated with 5-HT, adrenergic, and dopaminergic receptors. Surprisingly, 29100_SODIUM_CHANNEL_BLOCKER is on the list as well.

C_267519 has seven unique activity strings, but nearly all are closely related, having to do with CCK/gastrin. However, 78418_PHOSPHODIESTERASE_IV_INHIBITOR is an unexpected activity.

There are a number of substructures that are inhibitors of enzymes. C_194637 has activities all associated with metalloproteases. C_172433 is associated with phosphodiesterases, the inhibition of which is known to prevent the

Table 1. Unique Activities Associated with Consensus Substructures from the A. USPDI and B. MDDR[a]

A. USPDI

C_hydroxyprogesterone

ANTINEOPLASTIC	5	ANTI_INFLAMMATORY___STEROIDAL___OPHTHALMIC	1
ANABOLIC_STEROID	2	ANTI_INFLAMMATORY___STEROIDAL___RECTAL	1
ANDROGEN	2	CORTICOSTEROID___MINERALOCORTICOID	1
ANTIANEMIC	2	CORTICOSTEROID___OPHTHALMIC	1
ANTI_INFLAMMATORY___STEROIDAL	2	CORTICOSTEROID___RECTAL	1
CORTICOSTEROID	2	DIAGNOSTIC_AID___ESTROGEN_PRODUCTION	1
IMMUNOSUPPRESSANT	2	DIAGNOSTIC_AID___RENAL_TUBULAR_ACIDOSIS	1
ANDROGEN_ESTROGEN	1	ESTROGEN___SYSTEMIC	1
ANTIANGIOEDEMA___HEREDITARY___AGENT	1	OSTEOPOROSIS_PROPHYLACTIC	1
ANTIANORECTIC	1	OVARIAN_HORMONE_THERAPY_AGENT	1
ANTICACHECTIC	1	PROGESTATIONAL_AGENT	1
ANTIEMETIC___IN_CANCER_CHEMOTHERAPY	1	UROGENITAL_SYMPTOMS_SUPPRESSANT	1
ANTIHYPOTENSIVE___IDIOPATHIC_ORTHOSTATIC	1		

C_imipramine_HCl

ANTIDEPRESSANT	5	ANTIEMETIC	1
ANTINEURALGIC	5	ANTIENURETIC	1
ANTIPANIC_AGENT	4	ANTIHISTAMINIC___H_1_RECEPTOR	1
ANTIBULIMIC	2	ANTIOBSESSIVE_COMPULSIVE_AGENT	1
ANTICATAPLECTIC	2	ANTIPRURITIC	1
ANTIPSYCHOTIC	2	ANTIPRURITIC___TOPICAL	1
ANTIULCER_AGENT	2	SEDATIVE	1
ANESTHETIC_ADJUNCT	1	SEDATIVE_HYPNOTIC	1
ANTIDYSKINETIC___HUNTINGTON'S_CHOREA	1		

C_ferrous_citrate

NUTRITIONAL_SUPPLEMENT___MINERAL	2	ANTIUROLITHIC___CYSTINE_CALCULI	1
ALKALIZER___URINARY	1	ANTIUROLITHIC___URIC_ACID_CALCULI	1
ANTHELMINTIC___SYSTEMIC	1	DIAGNOSTIC_AID_RADIOACTIVE___NEOPLASTIC_ DISEASE	1
ANTIHYPERPHOSPHATEMIC	1		
ANTIHYPOCALCEMIC	1	DIAGNOSTIC_AID___IRON_ABSORPTION	1
ANTIHYPOMAGNESEMIC	1	DIAGNOSTIC_AID___IRON_METABOLISM	1
ANTIUROLITHIC___CALCIUM_PHOSPHATE_CALCULI	1	DIAGNOSTIC_AID___RADIOACTIVE___FOCAL_ INFLAMMATORY_LESIONS	1
ANTIUROLITHIC___CALCIUM_OXALATE_CALCULI	1		

C_alprostadil

DIAGNOSTIC_AID___ERECTILE_DYSFUNCTION	1	ANTIHEMORRHAGIC___POSTABORTION_UTERINE_ BLEEDING	1
DIAGNOSTIC_AID___PENILE_VASCULAR_IMAGING	1		
DUCTUS_ARTERIOSIS_PATENCY_ADJUNCT	1	OXYTOCIC	1
IMPOTENCE_THERAPY_AGENT	1	PROSTAGLANDIN	1
ANTIULCER_AGENT	1	DIAGNOSTIC_AID___ANGIOGRAPHY	1
GASTRIC_MUCOSA_PROTECTANT	1	UTERINE_STIMULANT	1
ABORTIFACIENT	1	ANTIHYPERTENSIVE___PULMONARY	1
ANTIHEMORRHAGIC___POSTPARTUM_UTERINE_ BLEEDING	1	VASODILATOR	1

C_meclocycline

ANTIPROTOZOAL	4	ANTIACNE_AGENT___TOPICAL	1
ANTIBACTERIAL_SYSTEMIC	4	ANTIMALARIAL	1
ANTIBACTERIAL_DENTAL	2	ANTIRHEUMATIC	1
ANTIBACTERIAL_TOPICAL	1	DIURETIC_SYNDROME_OF_INAPPROPRIATE...	1
ANTIBACTERIAL_OPTHALMIC	1	ENZYME_INHIBITOR___DENTAL	1
ANTIACNE_AGENT___SYSTEMIC	1	INTRAPLEURAL_SCLEROSING_AGENT	1

B. MDDR

C_265391

50050_TREATMENT_FOR_OSTEOPOROSIS	77	08000_ANTIDEPRESSANT	6
75000_ANTINEOPLASTIC	54	12342_ANTIMIGRAINE	6
33451_RESTENOSIS___AGENT_FOR	36	42731_SUBSTANCE_P_ANTAGONIST	6
41300_ESTROGEN_RECEPTOR_MODULATOR	32	50060_BONE_RESORPTION_INHIBITOR	4
75711_ANTIESTROGEN	32	51300_BONE_REGENERATION___AGENT_FOR	4
52000_HYPOLIPIDEMIC	28	40300_CONTRACEPTIVE	2
40210_ESTROGEN	24	30000_ANTIANGINAL	1
40340_POST_COITAL_CONTRACEPTIVE	8	37300_THROMBOLYTIC	1
01200_ANALGESIC___NON_OPIOID	6	40120_ANTIANDROGEN	1
02500_IMMUNOMODULATOR	6	41550_POSTMENOPAUSAL_SYNDROME___AGENT_FOR	1
06200_ANXIOLYTIC	6	80499_DIAGNOSTIC_FOR_CANCER	1

Table 1. (Continued)

		C_091258	
71000_ANTIVIRAL	57	72100_ANTIMALARIAL	2
75000_ANTINEOPLASTIC	39	72200_ANTITRICHOMONAL	2
78325_S_ADENOSYL_L_HOMOCYSTEINE_ HYDROLASE_INHIBITOR	25	72620_ANTITRYPANOSOMAL	2
		02150_DISEASE_MODIFYING_DRUG	1
71521_ANTIVIRAL___AIDS_	12	07707_ADENOSINE___A1___AGONIST	1
75100_ANTIMETABOLITE	9	28000_CARDIOTONIC	1
31000_ANTIHYPERTENSIVE	6	29000_ANTIARRHYTHMIC	1
72000_ANTIPROTOZOAL	6	30000_ANTIANGINAL	1
02000_ANTIARTHRITIC	4	31300_VASODILATOR	1
02100_ANTIINFLAMMATORY	4	55210_INFLAMMATORY_BOWEL_DISEASE___AGENT	1
33500_SEPTIC_SHOCK___TREATMENT_FOR	3	62000_IMMUNOMODULATOR	1
62200_IMMUNOSUPPRESSANT	3	71300_VIRAL_HEPATITIS___AGENT_FOR	1
		C_091258	
78385_ADENOSINE_KINASE_INHIBITOR	3	75310_PLATINUM_COMPLEX	1
02454_TNF_INHIBITOR	2	78330_PROTEASE_INHIBITOR	1
12340_MULTIPLE_SCLEROSIS___AGENT_FOR	2	78353_ADENOSINE_DEAMINASE_INHIBITOR	1
12452_NEURONAL_INJURY_INHIBITOR	2	78366_RIBONUCLEOTIDE_REDUCTASE_INHIBITOR	1
71522_REVERSE_TRANSCRIPTASE_INHIBITOR	2	80000_DIAGNOSTIC_AGENT	1
		C_144938	
75000_ANTINEOPLASTIC	40	36300_HEMATOPOIETIC	2
78374_PROTEIN_KINASE_C_INHIBITOR	40	59300_ANTIPSORIATIC	2
09200_COGNITION_DISORDERS___AGENT_FOR	16	75400_ANTINEOPLASTIC_ANTIBIOTIC	2
11100_ANTIPARKINSONIAN	16	75820_ANTINEOPLASTIC_ENHANCER	2
62200_IMMUNOSUPPRESSANT	16	78373_TOPOISOMERASE_INHIBITOR	2
31000_ANTIHYPERTENSIVE	14	37200_PLATELET_ANTIAGGREGATORY	1
12452_NEURONAL_INJURY_INHIBITOR	13	38111_ANTITHROMBOCYTOPENIC	1
34000_DIURETIC	11	67000_ANTIBIOTIC	1
02100_ANTIINFLAMMATORY	10	70000_ANTIFUNGAL	1
09250_NEUROTROPHIC_FACTOR	8	75845_CHEMOPROTECTIVE	1
10000_ANTICONVULSANT	7	77000_RADIOPROTECTOR	1
12335_AMYOTROPHIC_LATERAL_SCLEROSIS___AGENT	7	78370_TYROSINE_SPECIFIC_PROTEIN_KINASE_INHIBITOR	1
62000_IMMUNOMODULATOR	4		
		C_182014	
75000_ANTINEOPLASTIC	46	59300_ANTIPSORIATIC	3
75752_ALKYLPHOSPHOLIPID	26	78348_PHOSPHOLIPASE_A2_INHIBITOR	3
31000_ANTIHYPERTENSIVE	17	12452_NEURONAL_INJURY_INHIBITOR	2
71000_ANTIVIRAL	15	57500_PANCREAS_DISORDERS___AGENT_FOR	2
27261_PAF_ANTAGONIST	11	84900_PHARMACOLOGICAL_TOOL	2
71521_ANTIVIRAL___AIDS_	10	02000_ANTIARTHRITIC	1
70000_ANTIFUNGAL	8	12340_MULTIPLE_SCLEROSIS___AGENT_FOR	1
31610_PAF_ANALOG	6	71522_REVERSE_TRANSCRIPTASE_INHIBITOR	1
02100_ANTIINFLAMMATORY	5	73000_ANTHELMINTIC	1
72000_ANTIPROTOZOAL	4	75100_ANTIMETABOLITE	1
		C_184619	
75000_ANTINEOPLASTIC	19	02400_CORTICOSTEROID	2
75721_AROMATASE_INHIBITOR	11	40300_CONTRACEPTIVE	2
59500_ANTIACNE	10	59200_ANTIINFLAMMATORY___TOPICAL	2
31000_ANTIHYPERTENSIVE	9	02000_ANTIARTHRITIC	1
40120_ANTIANDROGEN	9	41100_ANTIINFERTILITY___FEMALE	1
40220_PROGESTIN	8	43100_ANTIDIABETIC	1
78335_STEROID___5ALPHA___REDUCTASE_INHIBITOR	8	43200_ANTIDIABETIC___SYMPTOMATIC	1
35560_PROSTATE_DISORDERS___AGENT_FOR	7	52000_HYPOLIPIDEMIC	1
41200_ANTIINFERTILITY___MALE	4	53000_ANTIOBESITY	1
34000_DIURETIC	3	59300_ANTIPSORIATIC	1
34300_ALDOSTERONE_ANTAGONIST	3	78446_CYTOCHROME_P450_OXIDASE_INHIBITOR	1
59813_HAIR_GROWTH_PROMOTER	3	82100_DRUG_DELIVERY_SYSTEM	1
		C_197150	
07000_ANTIPSYCHOTIC	29	01200_ANALGESIC___NON_OPIOID	1
06235_5_HT1A_AGONIST	19	06232_5_HT2_ANTAGONIST	1
06200_ANXIOLYTIC	17	06240_5_HT1A_ANTAGONIST	1
08000_ANTIDEPRESSANT	8	06248_5_HT2A_ANTAGONIST	1
11100_ANTIPARKINSONIAN	7	07701_DOPAMINE___D2___ANTAGONIST	1
06245_5_HT_REUPTAKE_INHIBITOR	3	09400_PSYCHOSEXUAL_DYSFUNCTION___AGENT_FOR	1
11125_DOPAMINE___D2___AGONIST	3	12342_ANTIMIGRAINE	1
17100_ANTICHOLINERGIC___OPHTHALMIC	3	27200_ANTIALLERGIC/ANTIASTHMATIC	1
31000_ANTIHYPERTENSIVE	3	27240_MEDIATOR_RELEASE_INHIBITOR	1
27300_ANTIHISTAMINIC	2		

Table 1. (Continued)

	C_267519		
42711_CCK_ANTAGONIST	29	43100_ANTIDIABETIC	2
42712_CCK_A_ANTAGONIST	19	53000_ANTIOBESITY	2
57500_PANCREAS_DISORDERS__AGENT_FOR	11	53001_ANOREXIGENIC	2
54120_ANTIULCERATIVE	4	58210_IRRITABLE_BOWEL_SYNDROME__AGENT	2
54110_ANTISECRETORY__GASTRIC	3	12350_ANTIEMETIC	1
09200_COGNITION_DISORDERS__AGENT_FOR	2	27100_BRONCHODILATOR	1
11100_ANTIPARKINSONIAN	2	27200_ANTIALLERGIC/ANTIASTHMATIC	1
42705_CCK_A_AGONIST	2	42714_GASTRIN_ANTAGONIST	1
42710_CCK_AGONIST	2	78418_PHOSPHODIESTERASE_IV_INHIBITOR	1
42713_CCK_B_ANTAGONIST	2		
	C_091325		
75000_ANTINEOPLASTIC	5	06248_5_HT2A_ANTAGONIST	1
02454_TNF_INHIBITOR	3	07000_ANTIPSYCHOTIC	1
27200_ANTIALLERGIC/ANTIASTHMATIC	3	08000_ANTIDEPRESSANT	1
33500_SEPTIC_SHOCK__TREATMENT_FOR	3	12200_SKELETAL_MUSCLE_RELAXANT	1
01200_ANALGESIC__NON_OPIOID	2	18320_ANTIALLERGIC__OPTHALMIC	1
02100_ANTIINFLAMMATORY	2	26105_RHINITIS__AGENT_FOR	1
12342_ANTIMIGRAINE	2	27240_MEDIATOR_RELEASE_INHIBITOR	1
37200_PLATELET_ANTIAGGRETORY	2	27300_ANTIHISTAMINIC	1
43100_ANTIDIABETIC	2	59220_ALLERGIC_SKIN_DISORDERS__AGENT	1
06232_5_HT2_ANTAGONIST	1		

[a] The number in the second column indicates the number of molecules with that activity in the set consisting of the source molecule plus its HSCS-neighbors.

synthesis of TNF (tumor necrosis factor). However, C_240471 inhibits enzymes as diverse as collagenase, protein kinase, and phosphodiesterase.

Given an idea of what are interesting substructures common among several activities, one can use a QSAR method, for instance trend vector analysis,[14] to try to find the particular features that distinguish one type of activity from another. (Note that this is not a true QSAR because we are not dealing with uniformly measured activities of a set of molecules but only the *claimed* activities.) The MDDR is better than USPDI for QSAR because it has a larger number of more diverse molecules. Four representative substructure/activity combinations are in Figure 6. Here we used the AP descriptor for the trend vector analysis, but other descriptors lead to the same conclusions. In example A we are looking at compounds containing a substructure similar to C_265391 and asking if there is a structural difference between compounds claimed as antirestenosis agents (those that prevent reblocking of a coronary artery after stenting) and those claimed as antineoplastics. Despite the fact that both activities may work by antagonizing estrogen receptors, there is a statistically discernible difference. Antirestenosis agents tend to have W = −CH2−. Antineoplastic agents tend to have Y = −S− and tend to have hydroxyls on one or both of the lower aromatic rings (as in the substructure C_249850 in Figure 3B).

In example B we are looking at tricyclics and asking about the difference between antihistamines and antidepressants. The largest difference is that antidepressants have two methyls on the basic amine (or *N*-methylpiperidine), while the antihistamines tend to have longer substitutents on the amine. Also, antihistamines tend to have a heterocycle as one of the aromatic rings and are more likely to contain a carboxylate.

In example C we are looking at the difference between aryl-piperazine compounds that bind to dopamine receptors vs 5-HT receptors. The dopamine ligands tend to have a phenyl or a halogen-substituted phenyl as the aromatic ring. The 5-HT ligands tend to have a six-membered heterocycle as the aromatic ring and often have ortho-methoxy substituents. Also 5-HT ligands tend to have indole at the end of the aliphatic chain.

In example D we are trying to distinguish glucagon receptor antagonists from similar compounds with different activities, some of which include protein kinase inhibition. Glucagon receptor antagonists have N in the position marked 1, a C at position 2, and usually have a Cl at position 3. In contrast the other activities have a C at position 1 and N at position 2.

DISCUSSION

We have devised a simple method to identify multiactivity substructures in drug-like databases. The method described here can be applied to subsets of molecules or subsets of activities. In this work we applied the method to both in vivo activities and receptor-specific activitities.

Our method has some commonality with the work of others[15−17] where a common substructure is parsed from sets of active compounds. In our case, however, we are looking for common substructures in compounds with different activities instead of compounds with a common activity. Our current procedure is much like that in our previous work,[10] where we compared sets of molecules with two different user-selected activities, except here we are extending the comparison to all possible different activities and using all the molecules in a database. It is worth emphasizing that we have taken special care to ensure that these substructures, although associated with multiple activities, are larger and thus more specific than the common ring systems or other moieties found frequently in drug-like molecules, such as those listed, for instance, by Bemis and Murcko.[18,19] Some of our multiactivity substructures could be considered privileged structures, depending on one's definition of "privileged".

With any data-mining exercise, the results depend critically on what questions are asked and how. In our case, the results

Table 2. Unique Activities Associated with Selected Consensus Substructures from the MDDR Where Only Mechanism-Related Activities Are Considered

		C_160805		
06246_5_HT1D_AGONIST	16	06245_5_HT_REUPTAKE_INHIBITOR		1
31261_ADRENERGIC___ALPHA1___BLOCKER	9	06248_5_HT2A_ANTAGONIST		1
07710_DOPAMINE___D4___ANTAGONIST	5	29100_SODIUM_CHANNEL_BLOCKER		1
12452_NEURONAL_INJURY_INHIBITOR	3	31260_ADRENERGIC___ALPHA___BLOCKER		1
06240_5_HT1A_ANTAGONIST	1			
		C_267519		
42711_CCK_ANTAGONIST	29	42713_CCK_B_ANTAGONIST		2
42712_CCK_A_ANTAGONIST	19	42714_GASTRIN_ANTAGONIST		1
42705_CCK_A_AGONIST	2	78418_PHOSPHODIESTERASE_IV_INHIBITOR		1
42710_CCK_AGONIST	2			
		C_197150		
06235_5_HT1A_AGONIST	19	06240_5_HT1A_ANTAGONIST		1
06245_5_HT_REUPTAKE_INHIBITOR	3	06248_5_HT2A_ANTAGONIST		1
11125_DOPAMINE___D2___AGONIST	3	07701_DOPAMINE___D2___ANTAGONIST		1
06232_5_HT2_ANTAGONIST	1	27240_MEDIATOR_RELEASE_INHIBITOR		1
		C_240471		
78351_LIPOXYGENASE_INHIBITOR	16	27220_LEUKOTRIENE_SYNTHESIS_INHIBITOR		1
78371_COLLAGENASE_INHIBITOR	10	78370_TYROSINE_SPECIFIC_PROTEIN_KINASE_INHIBITOR		1
12453_LIPID_PEROXIDATION_INHIBITOR	3	78417_PHOSPHODIESTERASE_III_INHIBITOR		1
78331_CYCLOOXYGENASE_INHIBITOR	3	78418_PHOSPHODIESTERASE_IV_INHIBITOR		1
78454_CYCLOOXYGENASE_2_INHIBITOR	3			
		C_115397		
11123_DOPAMINE_AGONIST	13	11125_DOPAMINE___D2___AGONIST		1
08010_ADRENOCEPTOR___ALPHA2___ANTAGONIST	7	31262_ADRENERGIC___ALPHA2___BLOCKER		1
07702_DOPAMINE___D1___ANTAGONIST	1	42610_PROLACTIN_SECRETION_INHIBITOR		1
07705_DOPAMINE_AUTORECEPTOR_AGONIST	1			
		C_194637		
78371_COLLAGENASE_INHIBITOR	16	34620_NEUTRAL_ENDOPEPTIDASE_INHIBITOR		2
01141_AMINOPEPTIDASE_INHIBITOR	2	50060_BONE_RESORPTION_INHIBITOR		2
02454_TNF_INHIBITOR	2	78432_MATRIX_METALLOPROTEINASE_INHIBITOR		1
31410_ACE_INHIBITOR	2			
		C_242842		
06240_5_HT1A_ANTAGONIST	4	07710_DOPAMINE___D4___ANTAGONIST		1
07701_DOPAMINE___D2___ANTAGONIST	4	07712_SIGMA_ANTAGONIST		1
12452_NEURONAL_INJURY_INHIBITOR	4	11123_DOPAMINE_AGONIST		1
09221_ACETYLCHOLINESTERASE_INHIBITOR	2	31500_CALCIUM_CHANNEL_BLOCKER		1
06235_5_HT1A_AGONIST	1	33460_CALMODULIN_INHIBITOR		1
07705_DOPAMINE_AUTORECEPTOR_AGONIST	1			
		C_216903		
06235_5_HT1A_AGONIST	13	07703_DOPAMINE___D3___ANTAGONIST		1
06240_5_HT1A_ANTAGONIST	4	07705_DOPAMINE_AUTORECEPTOR_AGONIST		1
09249_MUSCARINIC___M1___AGONIST	3	11123_DOPAMINE_AGONIST		1
06246_5_HT1D_AGONIST	1			
		C_172433		
78418_PHOSPHODIESTERASE_IV_INHIBITOR	19	27220_LEUKOTRIENE_SYNTHESIS_INHIBITOR		1
02454_TNF_INHIBITOR	10	28310_PHOSPHODIESTERASE_INHIBITOR		1
12452_NEURONAL_INJURY_INHIBITOR	1	78417_PHOSPHODIESTERASE_III_INHIBITOR		1
		C_205639		
12452_NEURONAL_INJURY_INHIBITOR	7	12457_AMPA_RECEPTOR_ANTAGONIST		2
12455_NMDA_RECEPTOR_ANTAGONIST	6	06235_5_HT1A_AGONIST		1
06248_5_HT2A_ANTAGONIST	3	07712_SIGMA_ANTAGONIST		1
06245_5_HT_REUPTAKE_INHIBITOR	2			
		C_264017		
02456_IL_8_INHIBITOR	13	02455_IL_6_INHIBITOR		8
02450_IL_1_INHIBITOR	9	43132_GLUCAGON_RECEPTOR_ANTAGONIST		8
02454_TNF_INHIBITOR	8	02457_TNF_ALPHA_RELEASE_INHIBITOR		1

obviously depend to some degree on the various cutoffs used in our procedure. We have biased the selection of neighbors in step 3, for instance, toward fairly close analogues. In principle, we could use a smaller or no similarity cutoff and let the Z-score requirement for HSCS in step 4 control which pairs of molecules were kept. As the similarity cutoff gets smaller there are more neighbors and more diversity in the neighbors, and therefore the resulting consensus substructures would appear smaller and more generic. Also, the number of unique activities associated with each substructure gets larger. For instance at a similarity cutoff of 0.6, the best ranked consensus substructure has 249 neighbors with 68 activities vs 134 and 22 at 0.7. At 0.6 new interesting classes occasionally appear. For instance, in the MDDR, peptides

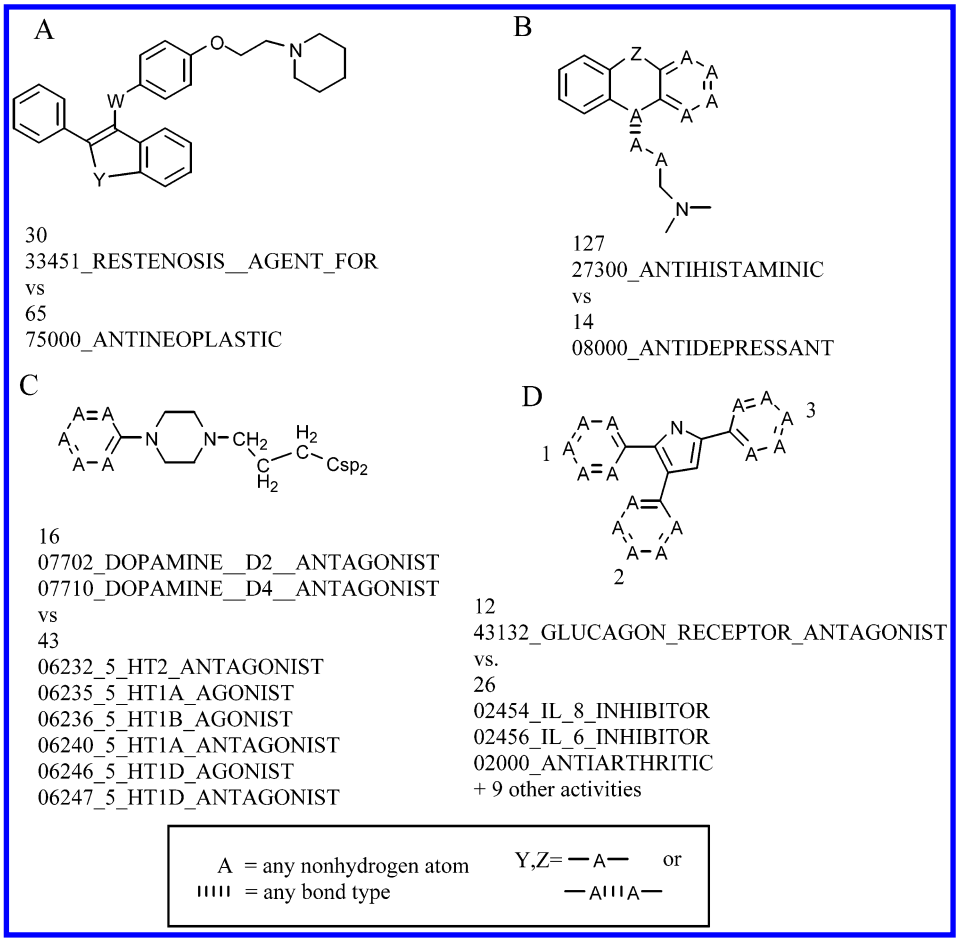


Figure 6. Example substructure/activity combinations used to find the structural features that distinguish one type of activity from another. Activities are from the MDDR database. In example A there are 30 compounds with 33451_RESTENOSIS__AGENT_FOR activity and 65 compounds with 75000_ANTINEOPLASTIC activity.

of the form Phe-X-Phe and Phe-Phe appear in the top 10. On the other hand, lowering the similarity cutoff greatly increases the computational cost and decreases the efficiency. For instance, at a cutoff of 0.7, there are ~68 000 pairs of compounds in the MDDR for which a HSCS must be calculated, of which ~55 000 (81%) are significant. At a cutoff of 0.6, there are ~880 000 pairs, of which ~327 000 (37%) are significant. Also, inspection of the neighbors at 0.6 shows that some of them would not be considered analogues by most chemists. Thus, a cutoff of 0.7 seems most useful at present.

Similarly, the drawing of the consensus substructure in step 5 depends on how much "conservation" is considered appropriate. We have assumed an atom in a molecule is conserved if it is shared among the majority (≥ 0.5) of the molecule's neighbors. This is consistent with our previous work, but other levels would also be reasonable depending on the use to which the substructure is to be put. Some examples are shown in Figure 7. Thus, substructures should not be taken unmodified from Figures 3 or 5 but need to be inspected and modified by a knowledgeable user before they are used as substructure search queries. For example, modifications to substructures were done for Figure 6.

There are many potential ways to sort the consensus substructures in step 6. For instance, sorting by the number of unique activities alone, or the ratio of unique activities to neighbors, are both reasonable. Finally, the elimination of redundant consensus substructures in step 7 can also be improved or replaced by a different algorithm.

The issues above can be addressed by algorithmic means. However, the major limitation for this technique is that it is very hard to tell a priori which activities are truly different. Ideally, one would like to have complete, consistent, mechanism-specific activity records for all molecules. However, current available databases that relate activities to molecules have several limitations that make them less than ideal. One important issue mentioned before is false negatives, in the sense that a compound C may not have activity A simply because it was never tested for activity A. Another limitation is that the databases are compiled from many different sources, and the criteria for a compound to be considered "active" may not be consistent from source to source, especially in regard to whether the activity was measured in vivo or in vitro. Then we have the important issue that the activity records may not be internally consistent, there may be synonymous records, the mechanisms of action are often not noted, or unknown, etc. Thus, a great deal of human interpretation needs to be done at the inspection phase, and the interpretation takes a fair amount of knowledge about the mechanisms by which various drugs act. We are working on an automatic way of determining whether activities are related. One approach for the MDDR is to use TIMI [20] to define a similarity between any two activity records based on whether the activity records are correlated with common

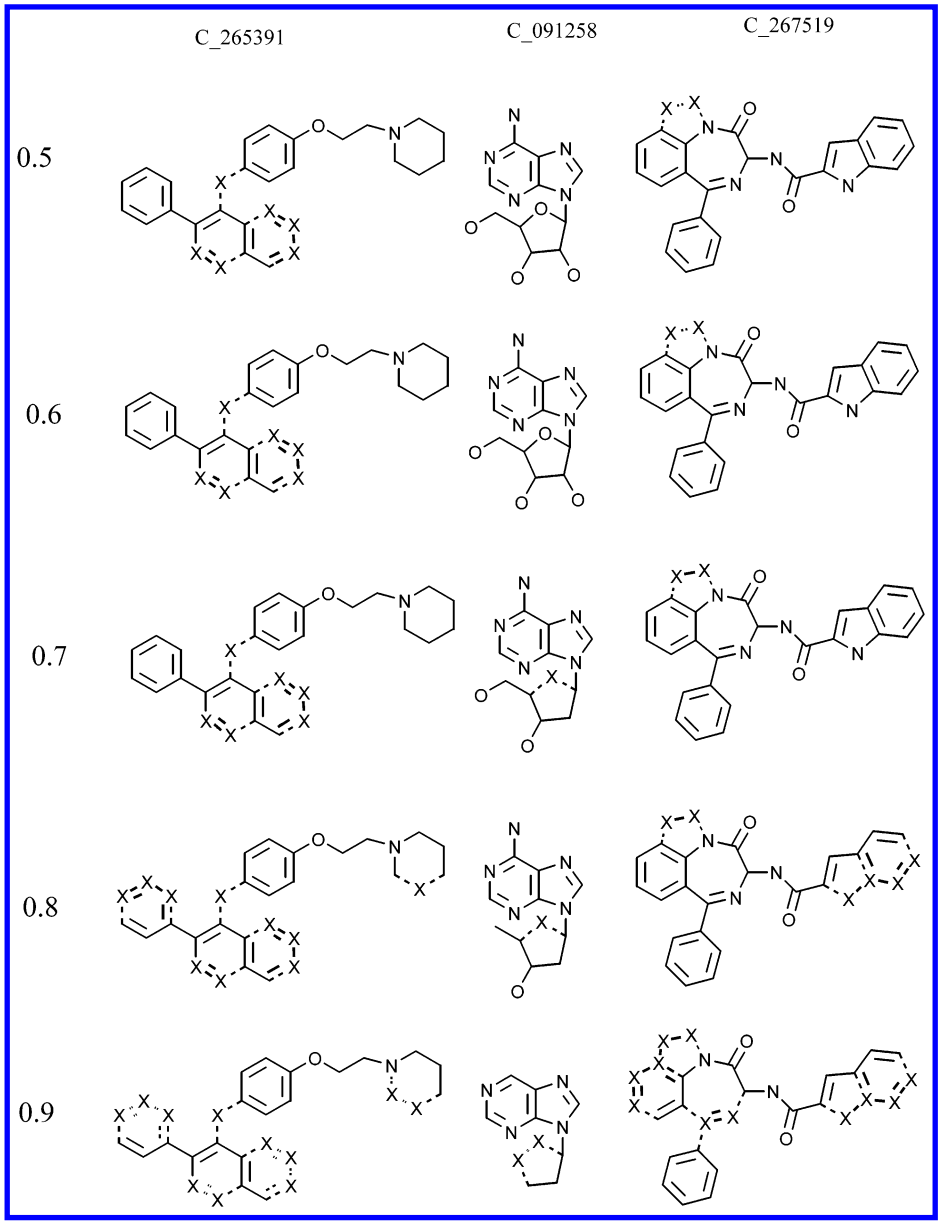


Figure 7. Selected consensus substructures from Figure 3B at different levels of "conservation". If an atom is in ≥0.5 of the neighbors of a molecule, it is considered conserved at 0.5, ≥0.6 conserved at 0.6, etc.

molecular descriptors and/or words from the MDDR "ACTION" fields and patent titles. That way if a set of activities, for example 42710_CCK_AGONIST, 42711_CCK_ANTAGONIST, 42705_CCK_A_AGONIST, and 42712_CCK_A_ANTAGONIST, were highly similar, each would be downweighted such that the set would count as only one unique activity instead of four.

Despite the limitations discussed above, we feel the preliminary results reported here are very encouraging in that they point out some known multiactivity substructures and suggest new ones. Data-mining can provide a rational basis for selecting or avoiding certain chemical classes. To take an obvious example, if one were designing a new antihistamine, one would certainly avoid tricyclics, for which both the USPDI and MDDR record a variety of CNS effects, unless one could establish molecular features that are strongly correlated with antihistamine effects and anticorrelated with CNS effects. One the other hand, if one were designing a combinatorial library to test specificity for a number of biogenic amine GPCR receptors, one might want to consider compounds of the form aryl-piperazine-$(CH2)_{3-4}$ because these have been claimed to bind to 5-HT, dopamine, and adrenergic receptors.

ACKNOWLEDGMENT

Dr. Simon K. Kearsley suggested a search for multiactivity substructures. Dr. Matt Walker made modifications to CKB that were useful in this project. The protocols for this work were written in the in-house modeling system MIX, and the author thanks the other members of the MIX team.

REFERENCES AND NOTES

(1) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing a privileged substructures. *J. Med. Chem.* **1999**, *42*, 2, 3251−3264.

(2) Nicolau, K. C.; Pfefferkorn, J. A.; Roecker, A. J.; Cao, G.-Q.; Barluenga, S.; Mitchell, H. J. Natural product-like combinatorial libraries based on privileged structures. 1. General principles and solid-phase synthesis of benzopyrans. *J. Am. Chem. Soc.* **2000**, *122*, 9939−9953.
(3) Horton, D. A.; Bourne, G. T.; Smythe, M. L. Exploring privileged structures: the combinatorial synthesis of cyclic peptides. *J. Comput.-Aided. Mol. Des*. **2002**, *16*, 415−430.
(4) Matter, H.; Baringhaus, K.-H.; Naumann, T.; Klabunde, T.; Pirard, B. Computational approaches toward the rational design of drug-like compound libraries. *Comb. Chem. High Throughput Screening* **2001**, *4*, 453−475.
(5) Patchett, A. A.; Nargund, R. P. Privileged structures—an update. *Annu. Rep. Med. Chem.* **2000**, *35*, 289−298.
(6) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discovery* **2002**, *1*, 882−894.
(7) USP DI Drug Information for the Healthcare Professional; Klasco, R. K., Ed.; MICROMEDEX, Inc.: Greenwood Village, CO, 2002. www.micromedex.com.
(8) MDDR Licensed by Molecular Design, Ltd. San Leandro, CA. www.mdli.com.
(9) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R.P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(10) Sheridan, R. P.; Miller, M. D. A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 915−924.
(11) Ho, T. K.; Hull, J. J.; Srihari, S. N. Decision combination in multiple classifier systems. *IEEE Trans Pattern Anal. Mach. Intel.* **1994**, *16*, 66−75.
(12) *Goodman and Gilman's Pharmacological Basis of Therapeutics*, 10th ed.; Hardman, J. G., Limbird, L. E., Goodman, A., Gilman, Eds.; McGraw-Hill: New York, 2001.
(13) Walker, M. J.; Hull, R. D.; Singh, S. B. CKB-the compound knowledge base: a text based chemical search system. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1293−1295.
(14) Sheridan, R. P.; Nachbar, R. B.; Bush, B. L. Extending the trend vector: the trend matrix and sample-based partial least squares. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 323−340.
(15) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. STIGMATA: an algorithm to determine structural commonalities in diverse subsets. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 862−871.
(16) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP−retrosynthetic combinatorial analysis procedure: a powerful new technique for indentifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 511−522.
(17) Lewell, X. Q.; Smith, R. Drug-motif-based diverse monomer selection: method and application in combinatorial chemistry. *J. Mol. Graph. Model*. **1997**, *15*, 43−48.
(18) Bemis, G.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(19) Bemis, G.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.
(20) Singh, S. B.; Hull, R. D.; Fluder, E. M. Text influenced molecular indexing (TIMI): a literature database that handles text and chemistry. *J. Chem. Inf. Comput. Sci*. In press.

CI030004Y