Properties of Known Drugs. 2. Side Chains

Guy W. Bemis* and Mark A. Murcko

Vertex Pharmaceuticals, 130 Waverly Street, Cambridge, Massachusetts 02139-4242

Received August 5, 1999

We continue our study of the common features present in drug molecules by looking in detail at drug side chains. Using shape description methods, we divide a database of commercially available drugs into a list of common drug side chains. On the basis of the atom pair shape descriptor (taking into account atom type, hybridization, and bond order), there are 1246 different side chains among the 5090 compounds analyzed. The average number of side chains per molecule is 4, and the average number of heavy atoms per side chain is 2. If we ignore the carbonyl side chain, then there are approximately 15 000 occurrences of side chains. Of these 15 000 approximately 11 000 are from the "top 20" group of side chains. This suggests that the diversity that side chains provide to drug molecules is quite low. We discuss ways that this work could be used to provide guidance for molecular design efforts.

Introduction

The analysis of structures of known drugs can provide valuable information. By drawing lessons from their structures we may both gain insight into drug discovery projects long since brought to a successful completion and provide guidance for new drug discovery programs. We have previously described such an analysis of the frameworks of drugs;[1] now we turn our attention to drug side chains—the acyclic arrays of atoms attached to frameworks.

This information is useful for a variety of purposes. It provides a basis set of side chains that can be used for almost any medicinal chemistry endeavor whether computational or experimental. It is reasonable to believe that these side chains will generally be synthetically accessible, metabolically well understood, and toxicologically benign.

For a data set of drugs we extracted information from the Comprehensive Medicinal Chemistry (CMC) database[2] which contains two-dimensional and predicted three-dimensional structures and important biochemical properties for known drugs. The CMC database has been developed from Pergammon's Comprehensive Medicinal Chemistry series.[3]

Methods

The version of the CMC database that we used for this work (v. 94.1) includes 6 990 compounds. However, many of these do not meet our criteria for various reasons, e.g., imaging agents, dental resins, and veterinary compounds. Thus, our first task was to identify and remove these compounds. We eliminated all compounds for which no therapeutic activity class was listed, as well as compounds which fell into any one of a number of undesirable therapeutic classes.[4,5] Additionally, we removed drugs for which only partial 2-dimensional structural information was available.

After this process, the CMC database had 5 090 remaining entries.

We started with the very simple definition for side chain atoms that we used in our previous work[1] which is defined as follows. After removing hydrogen atoms, atoms on the periphery of the molecule (i.e., those connected to only one other atom) are labeled as side chain atoms and removed from the molecule. This process is repeated iteratively until all atoms in the original molecule are removed (acyclic molecules) or until all remaining atoms are connected to two or more other atoms. This remaining group of atoms is defined as the molecular framework. Contiguous groups of side chain atoms from the original molecule are defined as side chains (Figure 1).

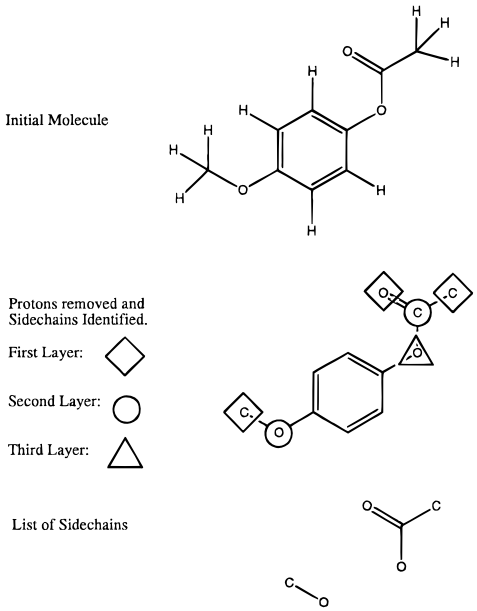


Figure 1. Example of the procedure for determination of side chains in a simple molecule.

When analyzing side chains in this manner it is important to recognize both side chain patterns and the way these patterns are connected to molecular frameworks. For instance, Figure 2 shows that the side chains from phenylacetic acid and phenyl acetate, when separated from their parent frameworks (Figure 2b), are identical (keep in mind that we are considering only heavy atoms). These side chains have different enough

10.1021/jm9903996 CCC: $18.00
Published on Web 12/01/1999

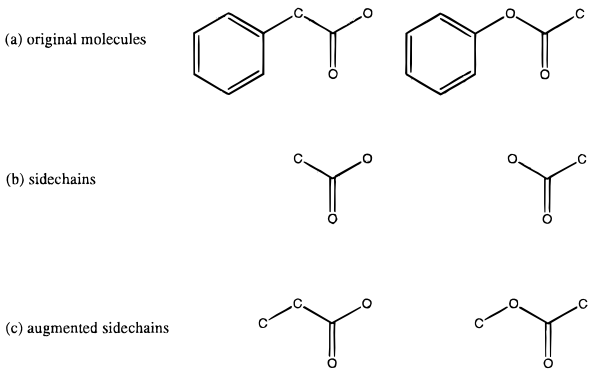


Figure 2. Example of side chain overlap problem. Side chains (b) determined from molecules (a) are identical. Augmented side chains (c) are necessary to distinguish these.

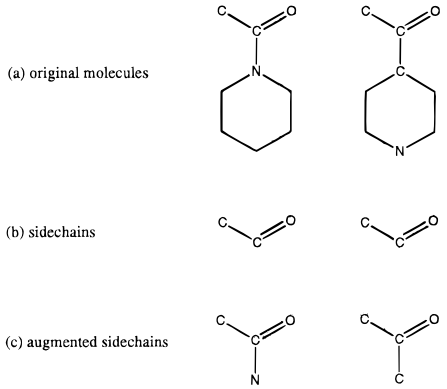


Figure 3. Example of an additional side chain overlap problem. Side chains (b) determined from molecules (a) are identical. Augmented side chains (c) are again necessary to distinguish these.

properties that any meaningful classification method should group these separately.

To address this concern, we include the framework atom directly connected to the side chain along with the side chain atoms. This provides a molecular moiety that can be usefully analyzed. We define this as an augmented side chain. Figure 2c shows augmented side chains for phenylacetic acid and phenyl acetate.

A more subtle difficulty in side chain categorization is provided by differences in framework atoms that are connected directly to the side chain. For instance, the acetyl side chain connected to the nitrogen of a piperidine framework has very different properties when compared to the acetyl side chain connected to one of the carbon atoms (Figure 3). Augmented side chains (Figure 3c) make this distinction.

A final modification to our definition is necessary because we need to retain information about the side chain atom that was originally the framework atom. Figure 4 shows the side chain classification scheme for methyl benzoate and phenyl acetate. The augmented side chains (Figure 4c) are unable to discriminate these groups without attachment of a "dummy" atom (Figure 4d) to the carbon atom derived from the framework. We define these side chain units to be "labeled side chains."

An additional consideration must be the suitability of our given molecular shape descriptor to represent the side chains as described above. We have found that one of the most effective molecular shape descriptors for small fragments is the atom pair descriptor.[6] Side

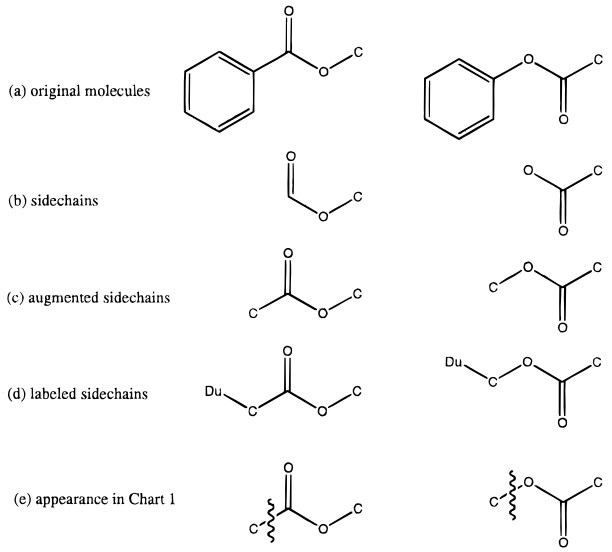


Figure 4. Example of augmented side chain overlap problem. Augmented side chains (c) derived from molecules (a) are identical. Only by inclusion of a dummy atom (Du) can the side chains be properly distinguished. Part e shows how results are displayed in Chart 1.

chains containing less than two heavy atoms, which we anticipated being well represented among drug molecules, cannot be described by atom pairs. Labeled side chains, which all contain two or more atoms, work correctly with atom pair descriptors.

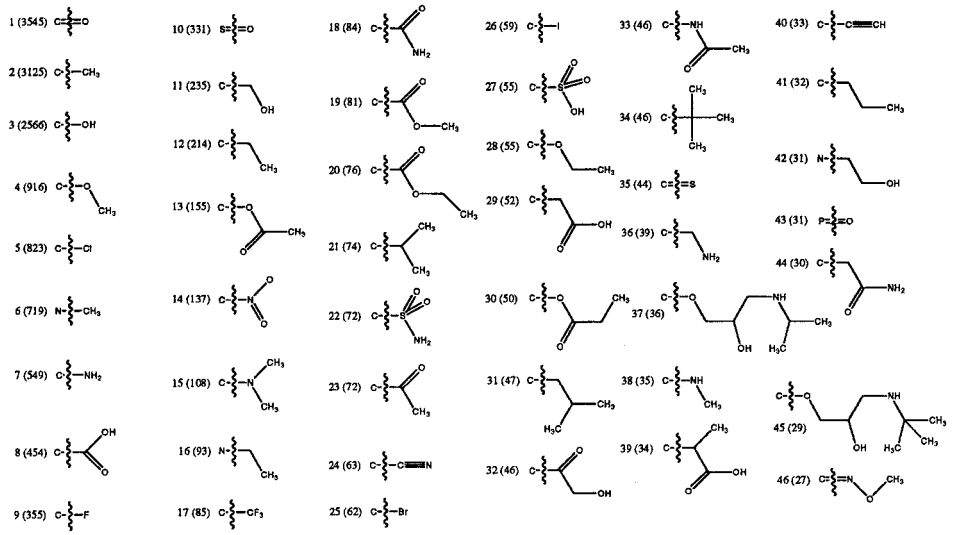
Results

Chart 1 shows the most frequently occurring drug side chains in our edited version of the CMC database. The drug side chains are represented as labeled side chains in this manner: the atom to the left of the wavy line is the framework atom to which the side chain is attached. The actual side chain atoms are to the right of the wavy line. Note that according to our original definition, carbonyl groups are considered to be side chains.

Out of 5090 drug molecules a total of 4689 contain labeled side chains. There are a total of 18 664 labeled side chains (including C═O) attached to these scaffolds. This means that the "average" drug scaffold contains four labeled side chains. Figure 5 shows a histogram for the number of labeled side chains per scaffold. It can be readily seen that most scaffolds have between one and five labeled side chains. Of those labeled side chains the vast majority (66%) contain one heavy atom in their actual side chain component (the number of heavy atoms to the right of the wavy line in Chart 1). Figure 6 is a histogram for the number of heavy atoms contained in each side chain. If we disregard X═O as a side chain, the one-atom side chains are reduced to 57%.

Additional information about side chains in drugs can be gained by looking at the frequency of occurrence for various pairs of labeled side chains (e.g., how many times are both a methyl side chain and a hydroxyl side chain found in a drug?). Table 1 shows the frequency of occurrence for pairs of the 25 most commonly occurring labeled side chains. Both the row number and column number correspond with the index numbers shown in Chart 1. Figure 7 shows a 3D graph for the top 10 labeled side chains in this data set. The most commonly found labeled side chain pairs from Figure 7 are

Chart 1. Labeled Side Chains from CMC[a]



[a] First number indicates rank order among drug side chains; number in parentheses is frequency of occurrence.

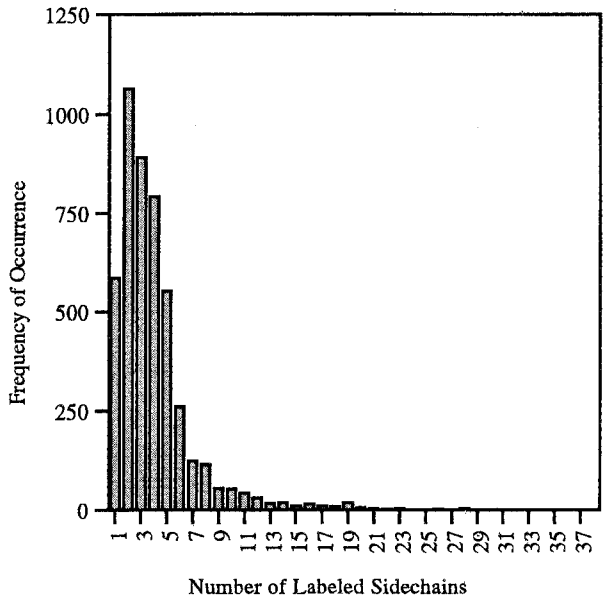


Figure 5. Histogram showing number of labeled side chains per drug framework.

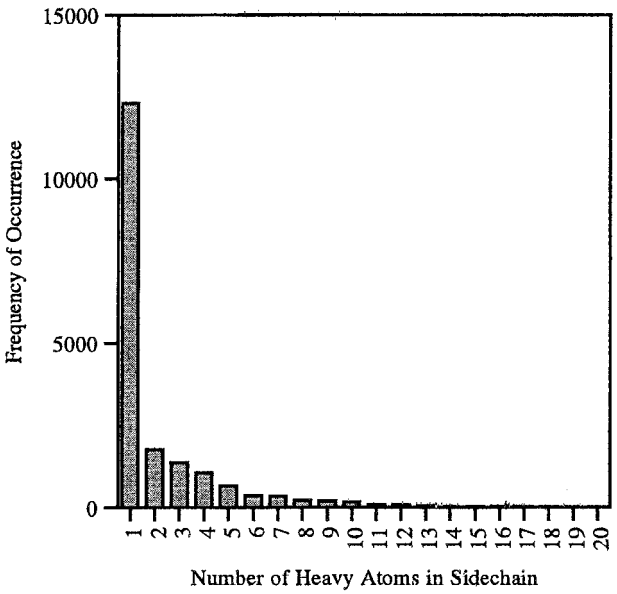


Figure 6. Histogram showing number of heavy atoms per side chain (number of atoms to the right of wavy line as shown in Chart 1).

enumerated at the top of Table 2. Also interesting to note are the labeled side chain combinations with the fewest occurrences. The least commonly found labeled side chain pairs from Figure 7 are enumerated at the top of Table 2. For comparisons, the distributions shown in Figure 7 and in Table 1 should ideally be normalized by the frequency of occurrence for each individual labeled side chain. This distribution should also be weighted by the expected number of labeled side chains in drug molecules (as shown in Figure 5) Determination of this normalization factor is quite complicated and will be the subject of future work.

Discussion

We have used a fairly simple scheme to classify drug side chains. When this work is combined with our previous drug framework classification work,[1] we have a comprehensive method for analyzing the frequency of occurrence of important atomic patterns in drug molecules.

To put our work into the context of other published methods for the analysis of molecular structures, we need to consider alternate classification schemes: bitmaps, shape descriptors, and biological methods.

Bitmaps represent molecules as lists of occurrences of discrete molecular features. These features are chosen to distinguish molecules based on their ability to represent the complete set of molecules being analyzed with a minimum of overlap. Bitmapped representations are most useful for tasks such as database searching for either whole molecules or substructures, molecular similarity calculations,[7] and information-based approaches to molecular property discriminants.[8] Cosgrove and Willett have used these methods to analyze molecules by their functional group content.[9]

Indeed, we could convert our framework and side chain work into bitmaps where one bit is set for the occurrence of each framework and each side chain. These bitmaps could prove to be very useful for molec-

Table 1. Labeled Side Chain Pair Distribution for Top 25 Labeled Side Chains[a]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	686																								
2	488	412																							
3	139	127	172																						
4	80	57	21	72																					
5	65	60	49	23	74																				
6	89	83	39	17	11	51																			
7	138	48	21	16	13	32	7																		
8	98	83	9	18	6	6	27	36																	
9	34	131	5	4	3	37	4	7	27																
10	60	66	38	4	20	3	1	4	3	26															
11	63	35	24	9	6	0	4	4	2	10	15														
12	15	6	0	5	3	1	2	2	0	0	0	2													
13	57	60	32	7	5	4	4	1	1	26	4	1	5												
14	10	3	3	2	6	5	9	5	0	1	2	0	0	2											
15	6	7	2	4	1	2	1	4	1	0	1	2	0	0	2										
16	9	20	1	8	1	1	3	2	2	3	0	2	14	0	0	0									
17	33	21	19	4	7	3	0	1	1	12	7	7	2	0	0	1	8								
18	14	4	2	6	3	0	0	1	1	0	0	0	0	1	0	0	5	5							
19	17	19	7	3	19	3	1	1	2	3	1	0	0	0	0	1	0	0	10						
20	9	4	5	12	2	2	2	0	0	1	0	0	0	0	0	0	0	0	0	2					
21	53	20	2	11	0	5	1	6	0	1	14	0	0	0	1	1	0	0	0	0	0				
22	11	8	6	4	8	1	3	1	3	0	0	0	0	0	0	0	1	0	6	0	0	0			
23	7	2	5	4	5	2	0	3	1	2	2	0	0	0	0	1	0	0	0	0	1	0	3		
24	2	12	0	1	1	0	1	1	3	1	1	1	0	0	0	0	0	0	0	0	0	1	0	16	
25	9	8	3	0	4	2	3	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	1	1	10

[a] Row and column numbers corrrespond to rank order for labeled side chains shown in Chart 1.

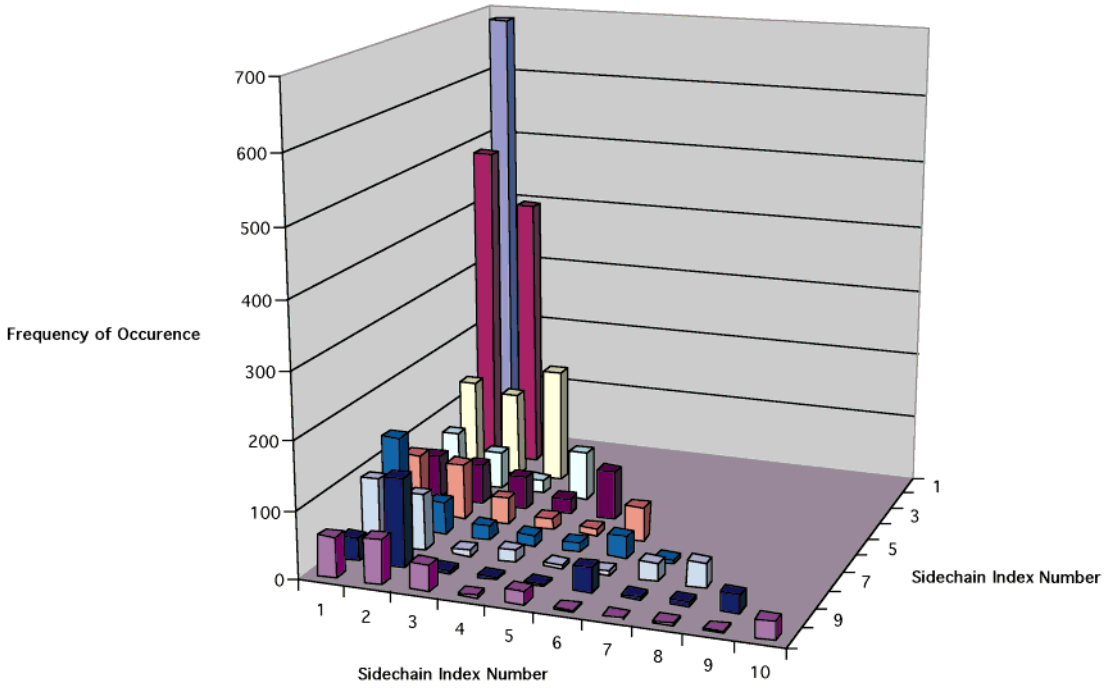


Figure 7. Labeled side chain pair distribution for the 10 most common labeled side chains. Side chain index number corresponds to rank order for labeled side chains shown in Chart 1.

ular discriminant determination, but the information lost, namely the pattern of side chain substitution around the framework, would make our current approach less useful for database structure retrieval schemes.

Shape descriptors represent molecules as either scalar numbers or vectors. By their nature they are representations of the entire molecule and therefore not suitable for identification of atom by atom description of the features found in drugs. Shape descriptors are most useful for tasks such as clustering molecules into groups that share similar overall shapes. Since molecular shape is related to boiling point, these descriptors have been used to particular advantage in prediction of the boiling points for hydrocarbons. Examples of molecular shape descriptors are topological torsions,[10] atom pairs,[6] Wiener indices,[11] and others.[7]

Our method of molecular analysis is unlikely to find use as a method of physical property prediction because we lose information about the placement of side chains relative to the molecular framework. This means we lose information about the gross shape features of molecules which are important for these predictions.

Biological methods classify molecules by their binding to a set of "receptor" proteins.[12,13] A bitmap-like representation is then constructed in which each "bit" rep-

Table 2. Most (top) and Least (bottom) Commonly Found Labeled Side Chain Pairs from Figure 7

side chain pair	frequency
Most Commonly Found	
C=O/C=O	686
C=O/C−CH_3	488
C−CH_3/C−CH_3	412
C−OH/C−OH	172
C=O/C−OH	139
C=O/C−NH_2	138
C−CH_3/C−F	131
C−CH_3/C−OH	127
C=O/C−CO_2H	98
C=O/N−CH_3	89
Least Commonly Found	
C−NH_2/S=O	1
C−Cl/C−F	3
N−CH_3/S=O	3
C−F/S=O	3
C−OCH_3/C−F	4
C−OCH_3/S=O	4
C−NH_2/C−F	4
C−CO_2H/S=O	4

resents the magnitude of binding to a specific protein. These methods may be considered a hybrid between bitmaps and shape descriptors. Attempts have also been made to perform these schemes computationally.[14] The usefulness of biological molecular classification schemes is still being determined.

An important characteristic of our work that is not found in the methods discussed above is the ability to synthesize new molecular structures. If our analysis is correct, that is, if we have identified features found in drug molecules, then the molecules that we generate are more likely to be "drug-like." We need merely start with a molecular framework, then based on the distributions shown in Figures 5 and 6 and Table 1 we graft side chains from Chart 1 to randomly chosen attachment points on the framework. Additionally, by including information about the points of attachment that originally connected the side chains to the frameworks, we have indirectly incorporated synthetic accessibility into our new molecular structures.

We could easily turn this into an iterative process in which a very large number of molecules could be generated. These molecules could be weeded out by a number of filters to ultimately derive a set of drug-like molecules optimized for a particular property prediction (e.g., predicted enzyme binding). We have used this approach to generate a combinatorial chemistry "friendly" set of molecules optimized for BBB penetration[15] and a generic set of minimally substituted expanded drug frameworks for docking studies.[16] One area of this work that could be addressed in the future is the absence of stereochemical information. This is a function of the atom-pair descriptors that we use for representation of the side chain atoms. We could incorporate a slightly more sophisticated molecular shape descriptor that accounts for stereochemistry.

Labeled side chains as proposed in this work are able to classify the appended functionalities of drug molecules into groups that correspond well with chemists' intuitive idea of what a side chain actually is and how a series of similar molecules should be grouped together. Previously published side chain schemes have concentrated mainly on peptide and peptide-like molecules.[17,18] The one slightly nonintuitive feature of our classification scheme is that "side chains" with cyclic groups are actually considered to be part of the molecular framework and not actually side chains. Once this is taken into consideration, our scheme becomes a powerful tool not only for classification of molecules but also for generation of new ones. Future work in this area could include using the results from Chart 1 to generate libraries of compounds with designed properties as well as using these results to refine more recent combinatorial chemistry based fragment libraries.[19,20]

Acknowledgment. We thank Ajay and Paul Charifson for comments on this manuscript.

References

(1) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(2) Comprehensive Medicinal Chemistry (CMC-3D) Release 94.1 is available from MDL Information Systems Inc., San Leandro, CA.
(3) Comprehensive Medicinal Chemisty, Vol. 6; Hansch, C., Sammes, P. G., Taylor, J. B., Series Eds.; Pergamon: Oxford, 1990.
(4) This is similar to a process used in the following: Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. Crystal Structure of HIV-1 Protease in Complex with VX-478, a Potent and Orally Bioavailable Inhibitor of the Enzyme. *J. Am. Chem. Soc.* **1995**, *117*, 1181−1182.
(5) Actual ISIS search terms used were the following: %veterinary%, %dental%, %radiopaque%, %contrast%, %solvent%, %local%, %insect%, %surfactant%, %sperm%, %aeros%, %wetting%, %flavor%, %al aids%, %ic aids%, %screen%, %emetic%, %topical%, %buffer%, %preserv%, %anesthetic%, %herbi%, %chelat.%, %keratoly%, %caries%, %dentist%, %adhesive%, %laxa%, %sweet%, %ecto%, %scabicide%, %astring%, %disinfectant%, %antiseptic%, %diagnos%, %fungi%, %plant%, %poultry%.
(6) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 82−85.
(7) Concepts and Applications of Molecular Similarity; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.
(8) Ajay, Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(9) Cosgrove D. A.; Willett P. SLASH: a program for analysing the functional groups in molecules. *J. Mol. Graph.* **1998**, *16*, 19−32.
(10) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(11) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(12) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.
(13) Kauvar, L. M.; Villar, H. O.; Sportsman, J. R.; Higgins, D. L.; Schmidt, D. E. Jr. Protein affinity map of chemical space. *J. Chromatogr. B. Biomed. Sci. Appl.* **1998**, *715*, 93−102.
(14) Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.
(15) Ajay; Bemis, G. W.; Murcko, M. A. Desigining Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.
(16) Charifson, P. S. Unpublished results.
(17) Fauchere, J.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Amino Acid Side Chain Parameters for Correlation Studies in Biology and Pharmacology. *Int. J. Pept. Protein Res.* **1988**, *32*, 269−278.
(18) Callantes, E. R.; Dunn, W. J., III. Amino Acid Side Chain Descriptors for Quantitative Structure−Activity Relationship Studies of Peptide Analogues. *J. Med. Chem.* **1995**, *38*, 2705−2713.
(19) Murray, C. W.; Clark, D. E.; Auton, T. R.; Firth M. A.; Li, J.; Sykes, R. A.; Waszkowycz, B.; Westhead, D. R.; Young, S. C. PRO_SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. *J. Comput. Aided Mol. Des.* **1997** *11* (2), 193−207.
(20) 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.

JM9903996