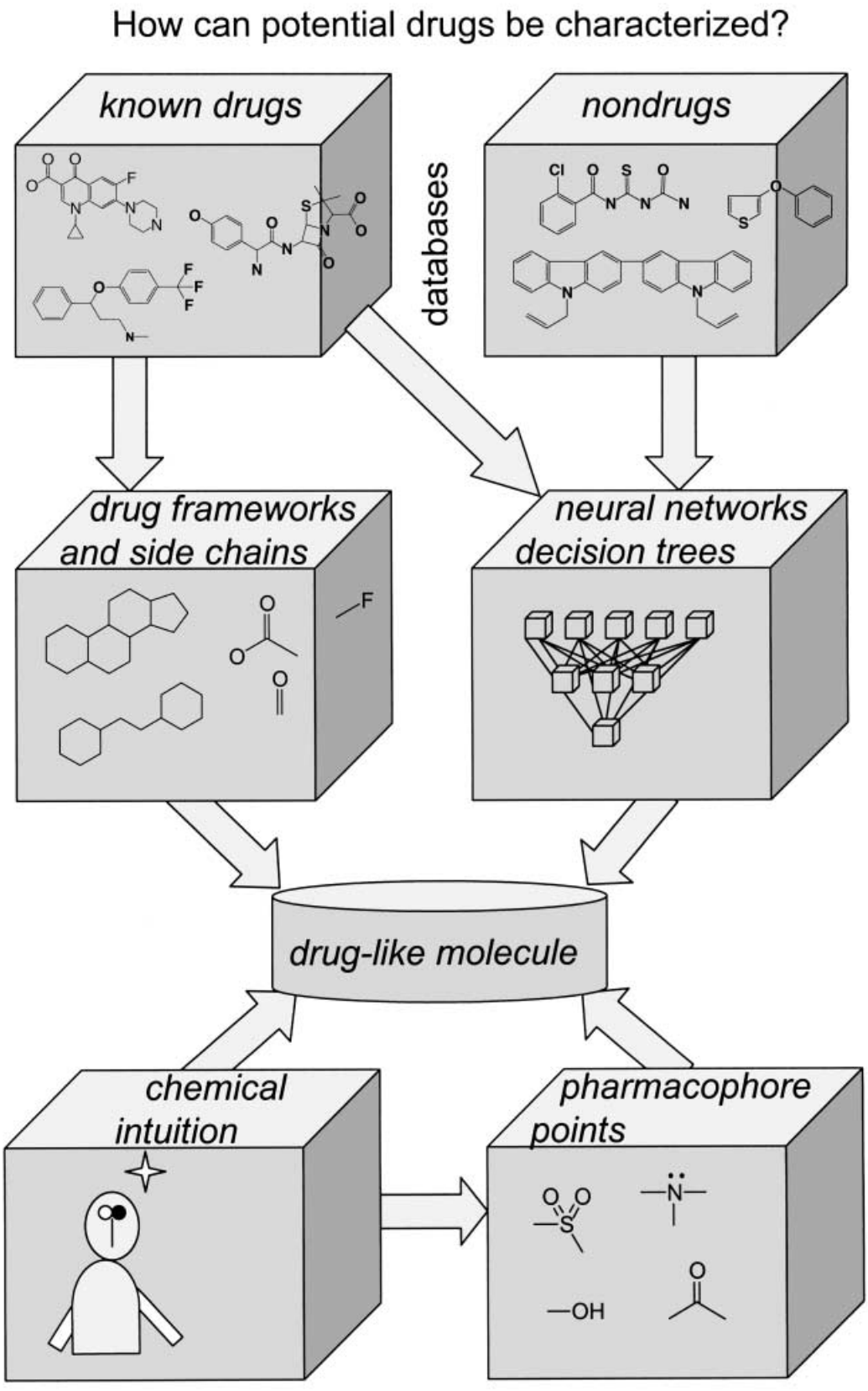


0947-6539/02/0809-1976 $ 20.00+.50/0 *Chem. Eur. J.* **2002**, *8*, No. 9

Pharmacophore Features of Potential Drugs

Ingo Muegge*[a]

Abstract: Drug discovery efforts rely increasingly on the identification of quality lead compounds through high-throughput synthesis and screening. However, large-scale random libraries have yielded only a low number of quality lead molecules. To address this shortcoming researchers have paid more attention to the concept of "drug-likeness" of molecules in combinatorial and screening libraries. Database profiling and analysis methods have been employed to identify the structural features of known drug molecules. Neural networks and machine learning methods help to distinguish between drugs and nondrugs. More recently, database-independent pharmacophore filters have been introduced that provide simple intuitive rules to classify potential drugs.

Keywords: computer chemistry · drug design · drug-like · high-throughput screening · library design

Introduction

With the advent of combinatorial chemistry[1, 2] and high-throughput screening (HTS)[3] finding leads in drug discovery has become a numbers game. Pharmaceutical companies typically screen close to one million compounds against tens of drug targets each year. Researchers in the early nineties believed that, with an increased number of compounds synthesized and screened, the number of drug candidates would increase in parallel. Consequently, early attempts in the design of combinatorial libraries focused on the synthesis of large random libraries that were sometimes optimized for diversity.[4] Unfortunately, the success rate of these early random libraries is considered to be low. The somewhat disappointing performance can be attributed, for example, to the high flexibility of synthesized molecules as well as to their high lipophilicity.[5] In addition, the libraries were often limited in their pharmacophore diversity.[6] It should also be noted here that, notwithstanding its successes, HTS produces a somewhat low number of viable lead compounds. For receptor and enzyme targets, on average one lead compound is identified for 120000 compounds screened.[7] HTS is much less successful for targets of protein – protein interaction.[7]

Due to the limitations of current HTS and library-design efforts, researchers have paid more attention to the concept of the "drug-likeness" of the molecules to be synthesized and screened. While the design of more drug-like libraries has often focused on profiling key physicochemical properties, such as molecular weight, charge, and lipophilicity,[5, 8] other researchers have focused on the specific functionalization of drug molecules.[9, 10] In this article, current concepts of the structural features of putative drug molecules are discussed. Table 1 gives an overview of current techniques for characterizing the structural features of drug-likeness and points out their advantages and disadvantages. For a broader view on the subject of drug-likeness we refer the reader to recent reviews.[11–14]

Structural Frameworks and Side Chains of Known Drugs

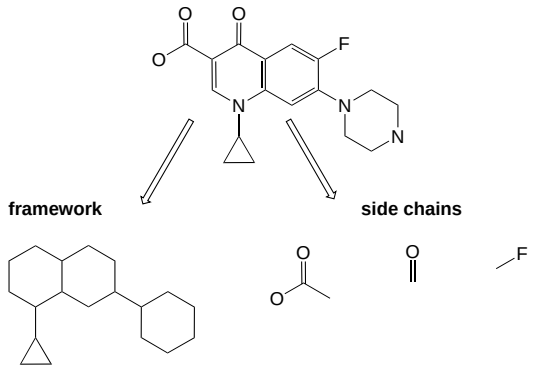
Bemis and Murcko have analyzed databases of commercially available drugs to identify common drug features using shape description methods.[15] They dissected drug molecules from the Comprehensive Medicinal Chemistry (CMC) database[16] into side chains and frameworks (containing ring systems and linkers). They found that only 32 frameworks are needed to describe the shapes of half the 5120 drugs in the CMC, which contains 1170 scaffolds. Scheme 1 shows the process of reducing a drug molecule to its framework, while Scheme 2 shows the most frequently occurring frameworks in the CMC. More recently, the side chains most abundant in drug molecules, have been analyzed by the same authors.[9] It has been found that of the 15000 side chains contained in the CMC, about 11000 belong to one of only 20 types of side chain including (starting with the most frequent): carbonyl, methyl, hydroxyl, methoxy, chloro, methylamine, primary amine, carboxylic acid, fluoro, sulfone. Most molecules possess between one and five side chains; the modal value being two side chains per molecule (in more than 20% of CMC). For the ten most frequently occurring side chains, Table 2 shows the most frequently occurring pairs of

[a] Dr. I. Muegge
Bayer Research Center, 400 Morgan Lane
West Haven, CT 06516 (USA)
Fax: (+1) 203-812-3505
E-mail: ingo.mugge.b@bayer.com

  0947-6539/02/0809-1977 $ 20.00+.50/0

Table 1. Current approaches to address the drug-likeness of compounds

Approach	Analysis	Drug/nondrug discrimination	Pros/Cons
Drug frameworks[15]	Database analysis of CMC	About half of all known drugs share only 32 common drug frameworks.	A limited number of frameworks (scaffolds) and side chains can be used to synthesize new molecular structures in combinatorial fashion that are similar to known drugs. However, drugs and nondrugs cannot be easily distinguished; e.g., benzene, the most frequently occurring framework in drugs, also occurs in 60% of reagent type compounds in the ACD.
Drug side chains[9]	Database analysis of CMC	11000 out of 15000 side chains of drugs in CMC belong to the "top 20" group of side chains.	
RECAP[28]	Fragmentation of molecules in drug databases around bonds formed by common chemical reactions.	Identifies privileged structural motifs in WDI correlated with biological activity against specific therapeutic classes.	De novo designed ligands are drug-like and more amenable to chemistry. Libraries focused on specific biological targets can be built. Fragments identified as privileged are biased by active analogues available in the database.
Neural networks[19, 20]	Chemical structure descriptors are used as input for neural nets trained on compound databases (CMC, MDDR, WDI, ACD)	Discriminates between drug databases and reagent databases with 65–90% accuracy.	High discrimination power between drugs and nondrugs. Classification is biased by databases used. Black box character of neural nets does not give structural classification rules.
Recursive partitioning[19, 21]	Chemical structure descriptors are used as input for decision trees trained on compound databases (WDI, ACD).	Classifies drugs/nondrugs with 70–83% accuracy.	High discrimination power between drugs and nondrugs. Simple rules can be derived that distinguish drugs from nondrugs. Classification is biased by databases used.
MLCC[23]	Up to tetracentered atom environments are assigned to molecules and their abundance in CMC and MDDR is tested.	76% of known drugs are classified correctly; only 19% of cancer drugs are recognized as drugs; it is suggested that current drug databases contain ≈80% of all viable drug types.	A general rule is used to describe any type of substructure. Disregarding the overall similarity of molecules, MLCC compares the local compatibility between molecules and known drugs. The method may be over-discriminative in that every group of a molecule has to be identical to some part of a known drug to have the molecule be classified as a drug.
Functional group analysis of drug database[29]	CMC is analyzed for the occurrence of functional groups	Benzene is most abundant (more than all heterocyclic rings combined); tertiary aliphatic amines, OH, and carboxamides are the most abundant functional groups.	Privileged functional groups occurring in drugs can be used for library design. The discrimination power between drugs and nondrugs is limited.
Drug-like index[30]	CMC is analyzed by using 25 structural descriptors and building blocks are clustered.	About 100 building blocks in CMC are statistically significant; 28% of ACD compounds are not drug-like.	The Drug-like index gives a quantitative measure of the drug-likeness of molecules, thereby ranking different molecules in a library for screening or combinatorial synthesis. It is biased by the databases (CMC) from which it is derived.
Pharmacophore point filter[10]	Intuitive rules are derived based on the observation that drug-like compounds need to be appropriately functionalized.	2/3 of CMC and MDDR and 1/3 of ACD are classified as drug-like.	Simple rules derived from "chemical intuition" provide moderate discrimination power between drugs and nondrugs. The method can easily be applied to library design because the rules apply to building blocks (rather than an enumerated molecule).



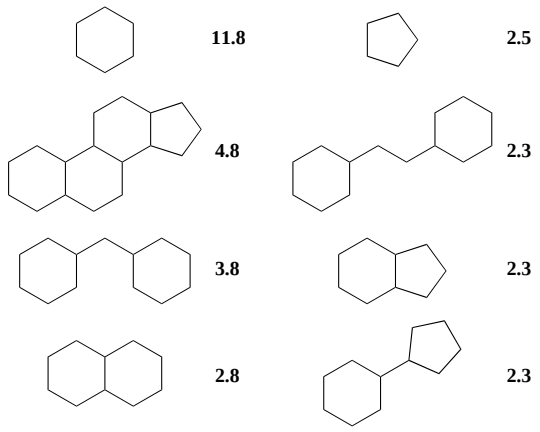
Scheme 1. Ciprofloxacin and its decomposition into framework and side chains.

side chains. It is interesting to note that six of these ten pairs contain two key polar groups, each of which is capable of forming hydrogen bonds and is also one of the key functional groups identified as pharmacophore points discussed below.

Drug–Nondrug Classification

The underlying assumption of topological drug classification schemes is that compounds that are structurally similar to known drug molecules may be potential drug candidates themselves—exhibiting desirable biological properties such as oral bioavailability, low toxicity, membrane permeability, metabolic stability, and reasonable clearance rates. Following this assumption, databases of drugs such as the CMC or MDDR[17] and other databases that presumably do not contain large numbers of drugs, for example reagent-like databases such as the ACD,[18] can be statistically analyzed to identify the structural features of molecules that help distinguish drugs from nondrugs. Drug-classification models that are based on this idea include neural network approaches[19, 20] as well as recursive partitioning approaches[19, 21] (Table 1). The nonlinear character of the neural network approaches prevents the derivation of discernible rules for the classification of compounds as drug or nondrug. The recursive partitioning approaches allow the structural

  0947-6539/02/0809-1978 $ 20.00+.50/0

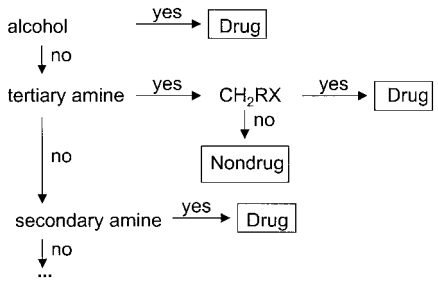


Scheme 2. The most frequently occurring frameworks in drugs. Numbers indicate percentages of occurrence in the CMC database.

Table 2. Most frequently found pairs of side chains in drugs in CMC.

Side chain pair	frequency [%]
C=O/C=O	14.6
C=O/C–CH_3	10.4
C–CH_3/C–CH_3	8.8
C–OH/C–OH	3.6
C=O/C–OH	2.9
C=O/C–NH_2	2.9
C–CH_3/C–F	2.8
C–CH_3/C–OH	2.7
C=O/C–CO_2H	2.1
C=O/N–CH_3	1.9

criteria for the classification of drugs and nondrugs to be established. Scheme 3 shows a portion of a decision tree derived from the WDI[22] and the ACD by Wagener and Geerestein. One example rule derived from this partial tree is: if a compound possesses no alcohol, does posses a tertiary aliphatic amine but not a methylene linker between a heteroatom and a carbon atom it is not drug-like. It is interesting to note that just by testing the presence or absence of hydroxyl, tertiary and secondary amines, carboxyl, phenol, or enol groups, 75 % of all drug-like structures in the MDDR and CMC can be recognized.



Scheme 3. Partial simplified decision tree of Wagener and Geerestein[21] for classifying compounds as drugs or nondrugs.

Pharmacophore Point Filter

The drug-fragmentation approaches combined with the machine-learning approaches discussed above suggest that the occurrence of a relatively small number of frameworks (ring structures and linkers), an even smaller number of side chains, and a few key polar groups characterize drug molecules very well. It has been observed that drugs distinguish themselves from nondrugs by possessing hydrophobic moieties that are well functionalized, while nondrugs often contain hydrophobic moieties that are underfunctionalized. Therefore, recent work focuses more on the presence of key functional groups in molecules.

A simple pharmacophore point filter has recently been introduced. It is based on the assumption that drug-like molecules should contain at least two distinct pharmacophore groups.[10] Four functional motifs have been identified that guarantee the hydrogen bonding capabilities that are essential for the specific interaction of a drug molecule with its biological target (Scheme 4). These motifs can be combined

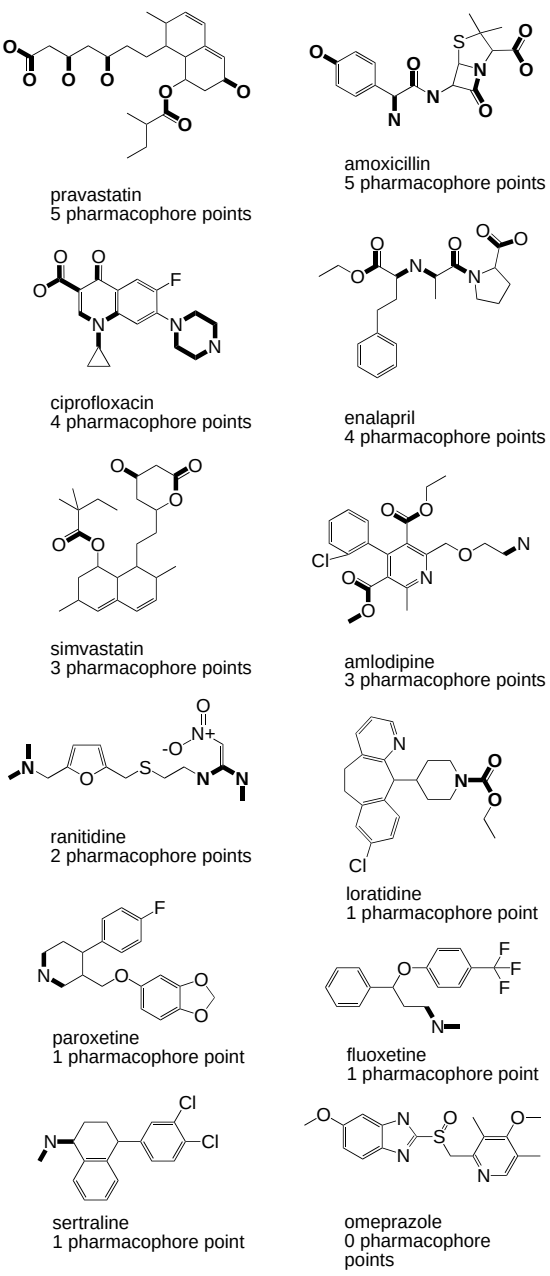
Scheme 4. Functional groups of drugs used to build pharmacophore points.

into functional groups, which are also referred to as pharmacophore points; they include: amine, amide, alcohol, ketone, sulfone, sulfonamide, carboxylic acid, carbamate, guanidine, amidine, urea, and ester. The following main rules apply to the pharmacophore point filter (PF1):

- Pharmacophore points are fused and counted as one if they are separated by less than two carbon atoms.
- Molecules with less than two and more than seven pharmacophore points fail the filter.
- Amines are considered pharmacophore points but not azoles or diazines.
- Compounds with more than one carboxylic acid are dismissed.
- Compounds without a ring structure are dismissed.
- Intracyclic amines in the same ring are fused to one pharmacophore point.

The requirement of two distinct pharmacophore points neglects at least one very important class of drugs—biogenic amine-containing CNS drugs. Therefore, a second pharmacophore filter has been designed that requires only one pharmacophore point in small molecules of the type amine, amidine, guanidine, or carboxylic acid (PF2).

Scheme 5 shows the pharmacophore count for the twelve top-selling drugs worldwide in 1997.[23] Seven of the twelve drugs possess between two and five pharmacophore points and pass PF1, four possess only one pharmacophore point, and one compound has none. Of the four compounds with only one pharmacophore point, three are serotonin re-uptake inhibitors that contain biogenic amines and cross the blood brain barrier (sertraline, fluoxetine, paroxetine). These compounds pass PF2. The classification shortcomings of the pharmacophore point filter are exemplified by the remaining two compounds. The H1 antihistamine loratidine, which does not cross the blood brain barrier,[24] has only one pharmacophore point. Omeprazole, a gastric acid secretion inhibitor,



Scheme 5. Pharmacophore points of the twelve top-selling drugs.

has no pharmacophore points. These two examples suggest that it may be desirable to allow for additional pharmacophore features such as sulfoxide or pyridine. The introduction of these two additional features into the pharmacophore point filter would allow loratidine and omeprazole to pass PF1. The power of PF1 to discriminate between drugs and nondrugs would not significantly change (the rate of filter survivors would increase by 3% in MDDR and 4% in ACD).

An analysis of drug databases and reagent-type databases reveal that about two thirds of drugs and nondrugs can be classified correctly by PF1. While the performance of this simple filter may not seem so impressive, the method has clear advantages over neural networks and decision trees. First, the rules are derived from "chemical wisdom". Drug databases certainly do not contain all feasible drugs, and therefore drug-like structural features are missing. In addition, reagent-like databases, used as negative controls, may very well contain drug-like compounds that are misclassified. In addition to this database bias, the neural net approach also has the shortcoming of being a "black box". That is, no direct rules for the design of drug-like molecules can be derived. This fact is particularly problematic when the drug-likeness of virtual combinatorial libraries needs to be assessed. The virtual library needs to be enumerated completely for this purpose. In contrast, the simple pharmacophore point filter is able to evaluate the drug-likeness of combinatorial libraries on the building-block level, thereby making the optimization process much more feasible.

Finally it should be noted that it is equally important to identify structural features in molecules that are unwanted in drugs, such as reactive or toxic moieties. An extensive survey of structural features in toxic chemicals in the RTECS database[25] reveals known carcinogens such as epoxy ethane or aromatic fused cycles as well as a large list of potentially toxic chemical frameworks.[26] A representative list of reactive fragments has been assembled, for example, by Rishton.[27]

[1] M. A. Gallop, R. W. Barrett, W. J. Dower, S. P. A. Fodor, A. M. Gordon, *J. Med. Chem.* **1994**, *37*, 1233 – 1251.
[2] E. M. Gordon, R. W. Barrett, W. J. Dower, S. P. A. Fodor, M. A. Gallop, *J. Med. Chem.* **1994**, *37*, 1385 – 1401.
[3] M. W. Lutz, J. A. Menius, T. D. Choi, R. G. Laskody, P. L. Domanico, A. S. Goetz, D. L. Saussy, *Drug Discovery Today* **1996**, *1*, 277 – 286.
[4] E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong, W. H. Moos, *J. Med. Chem.* **1995**, *38*, 1431 – 1436.
[5] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **1997**, *23*, 3 – 25.
[6] E. J. Martin, R. E. Critchlow, *J. Comb. Chem.* **1999**, *1*, 32 – 45.
[7] R. W. Spencer, *Biotechnol. Bioeng.* **1998**, *61*, 61 – 67.
[8] T. I. Oprea, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251 – 264.
[9] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1999**, *42*, 5095 – 5099.
[10] I. Muegge, D. Brittelli, S. L. Heald, *J. Med. Chem.* **2001**, *44*, 1841 – 1846.
[11] W. P. Walters, Ajay, M. A. Murcko, *Curr. Opin. Chem. Biol.* **1999**, *3*, 384 – 387.
[12] W. P. Walters, M. A. Murcko, *Methods Princ. Med. Chem.* **2000**, *10*, 15 – 30.
[13] D. E. Clark, S. D. Pickett, *Drug Discovery Today* **2000**, *5*, 49 – 58.
[14] B. L. Podlogar, I. Muegge, L. J. Brice, *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 102 – 109.
[15] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887 – 2893.
[16] Comprehensive Medicinal Chemistry is available from MDL Information Systems Inc., San Leandro, CA, 94577 and contains drugs already on the market.
[17] MACCS-II Drug Data Report is available from MDL Information Systems Inc., San Leandro, CA, 94577 and contains biologically active compounds in the early stages of drug development.
[18] Available Chemicals Directory is available from MDL Information Systems Inc., San Leandro, CA, 94577 and contains specialty bulk chemicals from commercial sources. Website www.mdli.com.
[19] Ajay, W. P. Walters, M. A. Murcko, *J. Med. Chem.* **1998**, *41*, 3314 – 3324.
[20] J. Sadowski, H. Kubinyi, *J. Med. Chem.* **1998**, *41*, 3325 – 3329.
[21] M. Wagener, V. J. van Geerestein, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280 – 292.
[22] World Drug Index is available from Derwent Information, London, UK. Website www.derwent.com.

  0947-6539/02/0809-1980 $ 20.00+.50/0 *Chem. Eur. J.* **2002**, *8*, No. 9

[23] J. Wang, K. Ramnarayan, *J. Comb. Chem.* **1999**, *1*, 524–533.
[24] A. M. Schleicher, *Acute Urticaria Emerg. Med.* **1998**, *30*, 143–145.
[25] RTECS C2(96–4); National Institute for Occupational Safety and Health (NIOSH), U.S. Department of Health and Human Services: Washington, D.C., **1996** URL: www.ccohs.ca.
[26] J. Wang, L. Lai, Y. Tang, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1173–1189.
[27] G. M. Rishton, *Drug Discovery Today* **1997**, *2*, 382–384.
[28] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
[29] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Comb. Chem.* **1999**, *1*, 55–68.
[30] J. Xu, J. Stevenson, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177–1187.

  0947-6539/02/0809-1981 $ 20.00+.50/0