Drug- and Lead-likeness, Target Class, and Molecular Diversity Analysis of 7.9 Million Commercially Available Organic Compounds Provided by 29 Suppliers

Alexander Chuprina,[†] Oleg Lukin,*[,†] Robert Demoiseaux,[‡] Alexander Buzko,[§] and Alexander Shivanyuk*[,#]

ChemBioCenter, National Taras Shevchenko University, 62, Volodymyrska Street, Kiev-33, 01033, Ukraine, Jonsson Comprehensive Cancer Center, University of California Los Angeles, 2805 Molecular Sciences Building, Los Angeles, California 90095-148, Abraxis Biosciences Incorporated, 11755 Wilshire Boulevard, 20th Floor, Los Angeles, California 90025, and The Institute of High Technologies, National Taras Shevchenko University, 62, Volodymyrska Street, Kiev-33, 01033, Ukraine

Received November 27, 2009

A database of 7.9 million compounds commercially available from 29 suppliers in 2008−2009 was assembled and analyzed. 5.2 million structures of this database were identified to be unique and were subjected to an assessment of physical and biological properties and estimation of molecular diversity. The rules of Lipinski and Veber were applied to the molecular weight, the calculated water/*n*-octanol partition coefficients (Clog *P*), the calculated aqueous solubility (log *S*), the numbers of hydrogen-bond donors and acceptors, and the calculated Caco-2 membrane permeability to identify the drug-like compounds, whereas the toxicity/reactivity filters were used to remove the structures with biologically undesired functional groups. This filtering resulted in 2.0 million (39%) structures perfectly suitable for high-throughput screening of biological activity. Modified filters applied to identify lead-like structures revealed that 16% of the unique compounds could be potential leads. Assessment of the biological activities, the analysis of diversity, and the sizes of exclusive sets of compounds are presented.

INTRODUCTION

High- and ultrahigh-throughput screening of large diverse libraries of organic compounds against multiple biochemical targets (enzymes, proteins, DNA, RNA, etc.) and living cells remains one of the most efficient and cost-effective approaches to finding hits in early drug discovery.[1] High demand of small molecule libraries spawned many chemical companies (mainly contract research organizations) specializing in combinatorial synthesis of low molecular weight organic compounds that might possess biological activities. Using methods of the automated and semiautomated high-throughput parallel synthesis, these companies produced millions of individual organic compounds whose structures are currently available online for structural search and purchase. Selection of organic compounds suitable for high-throughput screening of biological activity takes into account such parameters as physical properties, molecular geometry and presence of certain scaffolds, potential toxicity, and molecular diversity. In more rational selection procedures, docking studies (virtual screening) are carried out with the known molecular structure of a biological target. The simplest selection algorithms are based on applying sets of rigorous requirements to physical properties, and the presence of certain functional groups that are known to cause toxicity.[2–4] The description of the properties, including molecular weight, number of hydrogen-bond donors and acceptors, lipophilicity, and available polar surface area and the number of rotatable bonds was carefully validated by Lipinski[5] and Veber[6] and used in recent analyses of sizable collections of chemical compounds. Additional selection criteria include calculated water/*n*-octanol partition coefficient (log *P*),[7,8] solubility in water (log *S*),[9–12] and membrane permeability as well as the absence of chemical groups known to be reactive or cause toxicity. The series of tests performed on sets of reference compounds have revealed a good predictive power of the algorithms used for calculating logs *P* and *S*. Given the reliability of these calculations, it is now possible to perform quick analysis of ultra-large sets of compounds arriving at a collection of the drug- and lead-like structures. The analyses of commercially available collections of small organic molecules in terms of physical properties (drug- or lead-likeness) were carried out in 2004,[13] 2005,[14] and 2006.[15] These analyses used combined databases comprising 2.7, 3.8, and 5.3 million structures, respectively. Due to the rapid development of high-throughput screening and, hence, high-throughput parallel synthesis of drug- and lead-like compounds, the latest analysis of combined stock of commercially available screening compounds is apparently outdated. The present research has been undertaken in order to evaluate the suitability of the 2008−2009 combined stock of commercially available organic compounds for high-throughput screening of biological activity at early stages of drug discovery. In this work, we describe the analysis of 7.9 million organic compounds available from 29 commercial suppliers. Notably, in the pharmaceutical research, different

* Corresponding author. E-mail: oleg.lukin@univ.kiev.ua (O.L.); a.shivanyuk@univ.kiev.ua (A.S.).

[†] ChemBioCenter, National Taras Shevchenko University.

[‡] Jonsson Comprehensive Cancer Center, University of California Los Angeles.

[§] Abraxis Biosciences Incorporated.

[#] The Institute of High Technologies, National Taras Shevchenko University.

10.1021/ci900464s
Published on Web 03/18/2010

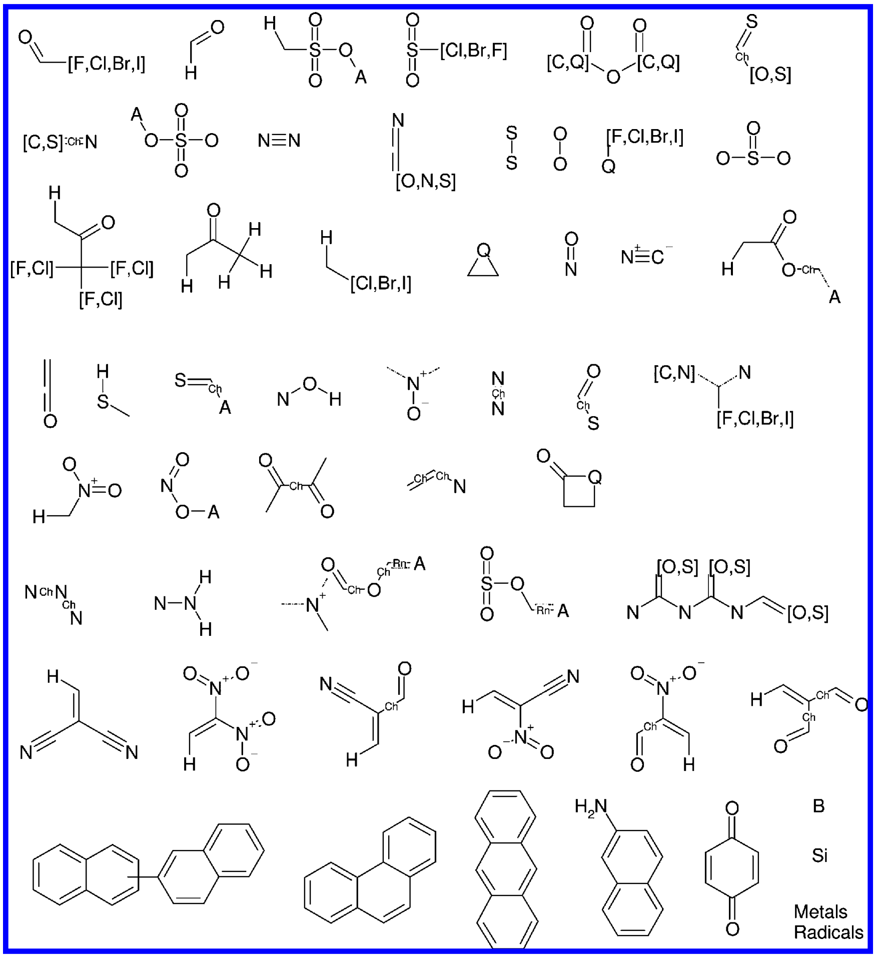

Figure 1. Toxicoforic and reactive structural fragments whose derivatives were removed during the selection of drug-like compounds. Rn and Ch stand for ring and open chain bonds, respectively.

cutoff values for calculated properties are applied to a database filtering dependently on a purpose. A pharmaceutical company usually provides vendors with precise property cutoff values to get compound sets fitting a concrete drug discovery project. To make our results comparable with those from the former database analyses, in this work, we use the traditional cutoff values (see Experimental Section) for calculated properties to identify drug- and lead-like compounds.

EXPERIMENTAL SECTION

Software. The MySQL database management system[16] was used to store structures, calculation parameters, and suppliers' data. The Simplified Molecular Input Line Entry System[17] (SMILES) representation was used to process the structures. The conversion of molecules into the SMILES format was carried out using the Jchem[18] program package utilities (molconvert, standardizer, cxcalc). Estimation of the physicochemical properties was performed with LigPrep and QikProp programs from the Schrödinger package.[19] The potential biological activity of the compounds was estimated and sorted with the aid of the Prediction of Activity Spectra for Substances, version 2005.5.1.7 (PASS) software.[20] The details of the activity assessment and the data sorting are given below in this section. Diversity coefficients were calculated with the CheD program.[21] The exclusion of compounds with toxic and reactive features shown in Figure 1 was performed utilizing an in-house written software linked to the structures via API MDL ISIS/Base.[22]

Library Preparation. The structures of commercially available organic compounds from 29 suppliers were acquired from MOLSOFT Web site[23] in secure digital (SD) format. The databases of individual suppliers were combined, and the structures were converted into the SMILES format. The SMILES representations were standardized by: (i) transforming nitro-, sulfoxide-, and nitroxide-groups into the charge-separated ones; (ii) transforming covalently represented alkali metal compounds into the ionic ones; (iii) removing saltdata from the structure field; (iv) removing the information on the absolute configuration of stereocenters; and (v) generating the canonic tautomers with Jchem. The duplicate structures were removed from the databases of individual suppliers, and the sets obtained were combined to give the collection of unique structures. The compounds provided by more than one provider were considered to be nonexclusive.

Selection of Drug- And Lead-Like Compounds. Property cutoff values used to select drug-like species are those of Lipinski and Veber; the values are given in the Supporting Information. The corresponding values used to select lead-like compounds are taken from a paper by Hann and Oprea[24] ($200 < MW < 460$, $-4 < \text{Clog } P < 4.2$, Hacc ≤ 9, Hdon ≤ 5, rotating bonds ≤ 10, PSA ≤ 170, CACO-2 ≥ 100, $-5 < \log S < 0.5$, absence of both toxic and reactive fragments).

Biological Activity Prediction. The PASS[20] software used in this study is capable of predicting of more than 3000 different types of biological activity with an average error

Table 1. Summary of Suppliers of Which Compound Libraries Were Used in This Work

supplier name	date of library	library size	exclusive compounds	% exclusive	% of all exclusive	address
Albany Molecular Research	2008.10	196 064	194 109	99.0%	4.9%	http://www.amriglobal.com
ART-CHEM	2008.06	110 873	25 124	22.7%	0.6%	http://www.art-chem.com
Asinex	2008.09	457 842	238 236	52.0%	6.0%	http://www.asinex.com
Asis Chem	2008.10	32 749	8976	27.4%	0.2%	http://www.asischem.com
ChemBridge	2009.03	741 176	321 033	43.3%	8.0%	http://www.chembridge.com
ChemDiv	2008.10	785 740	407 920	51.9%	10.2%	http://www.chemdiv.com
ChemStar	2008.06	28 946	7181	24.8%	0.2%	http://www.chemstar.ru
Enamine	2009.03	1 221 957	894 588	73.2%	22.4%	http://www.enamine.net
FluoroChem	2008.10	23 498	5760	24.5%	0.1%	http://www.fluorochem.net
InterBioScreen	2008.10	466 671	162 509	34.8%	4.1%	http://www.ibscreen.com
IVK Laboratories	2008.06	46 515	9544	20.5%	0.2%	http://www.ivklabs.com
Key Organics	2008.10	47 656	38 880	81.6%	1.0%	http://www.keyorganics.ltd.uk
Life Chemicals	2008.10	426 135	274 590	64.4%	6.9%	http://www.lifechemicals.com
Maybridge	2008.10	69 862	50 314	72.0%	1.3%	http://www.maybridge.com
Nanosyn	2008.10	62 597	10 984	17.5%	0.3%	http://www.nanosyn.com
Oakwood Chemicals	2008.10	12 621	928	7.4%	0.0%	http://www.oakwoodchemical.com
Otava Chemicals	2008.10	173 941	66 266	38.1%	1.7%	http://www.otavachemicals.com
Peakdale	2008.10	14 576	14 398	98.8%	0.4%	http://www.peakdale.co.uk
Pharmeks	2008.10	155 800	5494	3.5%	0.1%	http://www.pharmeks.com
Princeton Biomolecular Research	2008.10	380 424	40 191	10.6%	1.0%	http://www.princetonbio.com
SALOR	2008.06	48 693	22 407	46.0%	0.6%	http://www.sigmaaldrich.com
Specs	2008.10	223 630	60 710	27.1%	1.5%	http://www.specs.net
Spectrum	2008.10	8497	132	1.6%	0.0%	http://www.spectrum.kiev.ua
TimTec	2008.06	674 773	408 778	60.6%	10.2%	http://www.timtec.net
TOSLab	2008.10	26 713	10 399	38.9%	0.3%	http://www.toslab.com
Tripos	2008.10	154 604	135 216	87.5%	3.4%	http://www.tripos.com
Ufark	2008.10	28 881	21 418	74.2%	0.5%	http://ufark12.chem.ufl.edu
UORSY	2009.03	794 997	421 091	53.0%	10.5%	http://www.uorsy.com
Vitas-M Lab	2008.10	442 971	138 345	31.2%	3.5%	http://www.vitasmlab.com
total		7 859 402	3 995 521	50.8%		
total unique		5 183 506	3 995 521	77.1%		
open NCI database	2008.10	231 458	200 990	86.8%		http://cactus.nci.nih.gov

of 13%. The prediction algorithm is based on a compound similarity search in which the analyzed compound is compared to a series of validated biologically active species located in the program SAR database (60 722 compounds). The representation of all chemical structures for this analysis is done on the basis of the multilevel neighborhood of atoms (MNA) descriptors.[25] The probability for the compound to exhibit the given activity (P_a) and to be inactive (P_i) are then computed and compared. In this work, the following criteria of the potential biological activity were applied: $P_a > P_i$ and $P_a \geq 0.75$. The results were sorted into the following general types of biological activity: antiviral, G-protein-coupled receptor (GPCR), ion channel, kinase, and protease. A detailed list of the PASS-implemented activities was assigned to each of the general ones. The activities of the drug-like filtered data set comprising 2 020 027 unique structures were analyzed.

Fragment-Selected Classification. All unique cyclic (heterocycles and benzene) fragments were marked in the combined database with in-house written software. Twenty-five of the most frequently found fragments were analyzed with regard to their supplier-based distribution.

Diversity Calculations. Cluster analysis was performed using the ‘sphere exclusion’[26] method implemented in the JChem chemical fingerprint software. The similarity threshold was set to 0.8.

RESULTS

The results of the 29 supplier database analysis are compiled in Tables 1−4 and in Figures 2−8. For comparison purposes, the results of the parallel analysis of the National Cancer Institute (NCI) open database are provided. First, the tables and the figures will be described in general terms and then addressed in more detail. Table 1 lists the general information on suppliers whose compound catalogues were used in this work. Detailed results of applying different drug-like filters to the whole library of compounds from the 29 suppliers are given in Table 2 and Figures 2 and 3. Table 3 presents the results of applying lead-like filters to the combined database. Assessed biological activities of the library compounds and the exclusivity analysis of the offered biologically potent species are given in a form of three-dimensional (3D) diagrams in Figures 5 and 6, respectively.

Structural character of the suppliers’ compounds sorted by 25 cyclic fragments that are most frequently found in the combined database is shown in Figure 7. Finally, Table 4 and Figure 8 present evaluation of molecular diversity of the database compounds. Although the details of the analyses for each individual supplier are provided, mainly, general trends will be discussed.

The duplicate analysis showed that there are 5 183 506 unique structures constituting 66% of the total of 7 859 402 compounds in the database. The percentage of the unique structures in a combined database has shown a 5% increase since the 2004 analysis of Baurin et al.[13] This seems to be an indication that some companies have been focusing on increasing the number of exclusive compounds in their catalogues. Table 1 represents the database in the light of how many of the substances are exclusive to a particular supplier and the relative contributions of individual suppliers to the total amount of exclusive compounds. Suppliers with most sized libraries (>0.4 million) have 35−73% of exclusive compounds in their catalogues. These companies contribute 82% to the total database of exclusive compounds. An

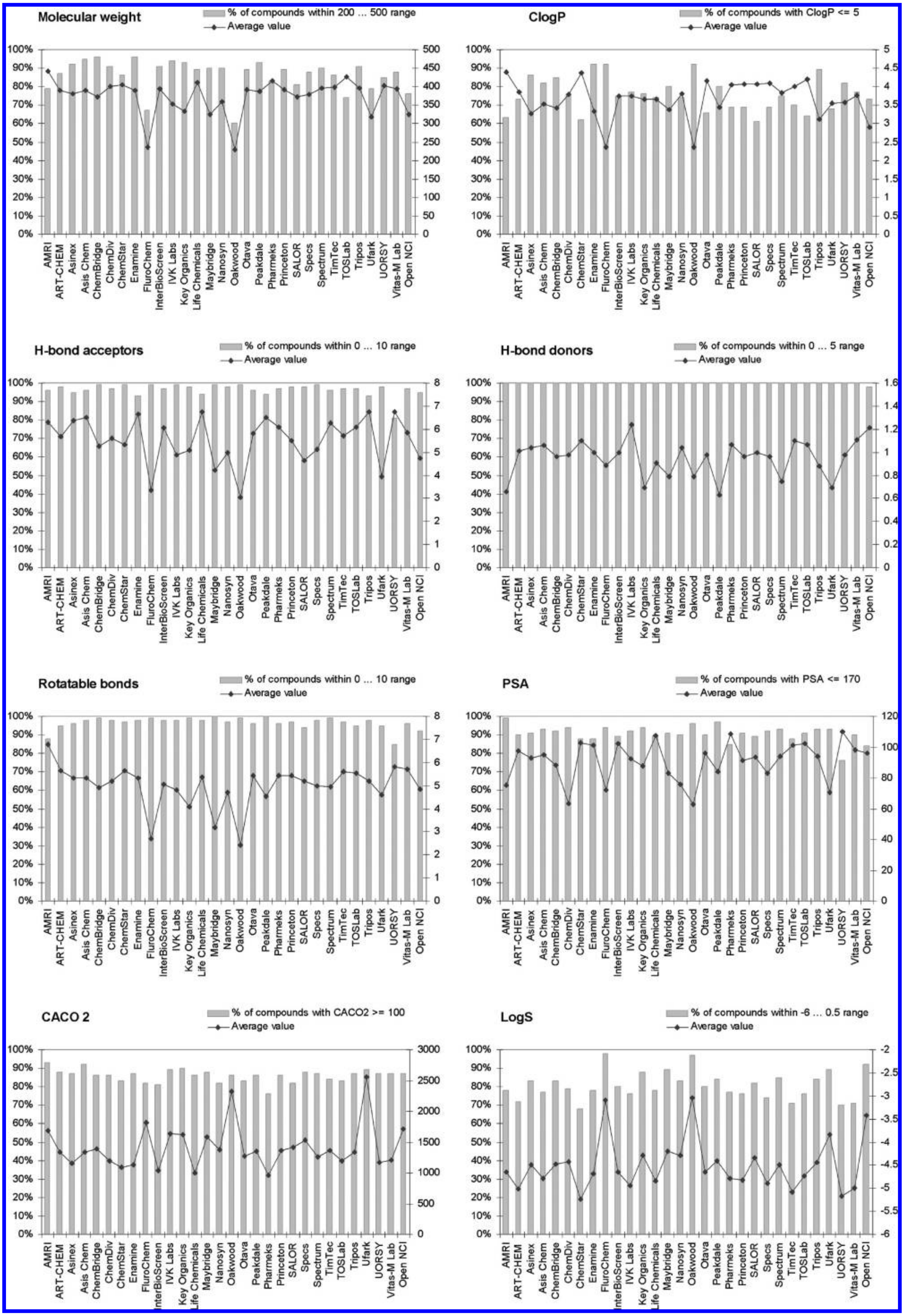

Figure 2. The results of applying of eight drug-like filters to compounds from 29 suppliers.

Table 2. Analysis of the Drug-like Properties of Compounds from 29 Suppliers

		3 of 4 Lipinski		4 of 4 Lipinski		Veber		absence of toxic/ reactive fragments		total filtered drug-like	
supplier	library size	No. of compounds passed	%	No. of compounds passed	%	No. of compounds passed	%	No. of compounds passed	%	No. of compounds passed	%
Albany Molecular Research	196 064	168 512	86%	104 334	53%	171 617	88%	186 901	95%	66 780	34%
ART-CHEM	110 873	101 959	92%	73 540	66%	94 653	85%	69 628	63%	35 040	32%
Asinex	457 842	443 550	97%	350 843	77%	399 447	87%	378 697	83%	212 472	46%
Asis Chem	32 749	31 834	97%	24 827	76%	30 118	92%	24 491	75%	14 031	43%
ChemBridge	741 176	728 794	98%	600 477	81%	678 668	92%	607 613	82%	386 566	52%
ChemDiv	785 740	747 052	95%	540 565	69%	720 162	92%	650 781	83%	330 144	42%
ChemStar	28 946	26 030	90%	16 336	56%	24 833	86%	13 288	46%	4977	17%
Enamine	1221 957	1160 859	95%	1 002 004	82%	1 063 102	87%	1 026 443	84%	672 076	55%
FluoroChem	23 498	23 040	98%	14 150	60%	22 027	94%	15 744	67%	7435	32%
InterBioScreen	466 671	444 365	95%	317 429	68%	405 300	87%	324 547	70%	144 850	31%
IVK Laboratories	46 515	45 144	97%	34 059	73%	41 926	90%	28 966	62%	13 688	29%
Key Organics	47 656	46 301	97%	33 383	70%	44 393	93%	33 065	69%	17 984	38%
Life Chemicals	426 135	401 386	94%	269 897	63%	372 876	88%	350 943	82%	153 850	36%
Maybridge	69 862	68 088	97%	48 988	70%	63 683	91%	46 238	66%	25 225	36%
Nanosyn	62 597	59 158	95%	42 241	67%	54 502	87%	37 063	59%	17 924	29%
Oakwood Chemicals	12 621	12 308	98%	6715	53%	12 046	95%	8500	67%	3796	30%
Otava Chemicals	173 941	159 692	92%	103 736	60%	150 082	86%	122 534	70%	49 281	28%
Peakdale	14 576	14 072	97%	10 246	70%	14 009	96%	13 903	95%	7209	49%
Pharmeks	155 800	138 626	89%	94 023	60%	127 982	82%	95 591	61%	34 811	22%
Princeton Biomolecular Research	380 424	352 728	93%	240 877	63%	335 530	88%	246 944	65%	107 968	28%
SALOR	48 693	44 053	90%	24 386	50%	40 715	84%	21 846	45%	8300	17%
Specs	223 630	207 769	93%	140 776	63%	201 236	90%	159 595	71%	71 882	32%
Spectrum	8497	7997	94%	5607	66%	7769	91%	7083	83%	3590	42%
TimTec	674 773	609 034	90%	431 965	64%	579 690	86%	390 235	58%	151 164	22%
TOSLab	26 713	22 748	85%	13 505	51%	23 045	86%	17 236	65%	5212	20%
Tripos	154 604	150 949	98%	115 517	75%	141 993	92%	129 010	83%	72 350	47%
Ufark	28 881	26 975	93%	14 984	52%	25 636	89%	15 813	55%	7650	26%
UORSY	794 997	683 318	86%	596 784	75%	591 882	74%	532 054	67%	277 088	35%
Vitas-M Lab	442 971	419 064	95%	302 404	68%	382 768	86%	327 505	74%	143 146	32%
unique	5 183 506	4 909 717	95%	3 716 537	72%	4 545 593	88%	4 010 470	77%	2 020 027	39%
open NCI database compounds	231 458	210 856	91%	161 543	70%	182 841	79%	140 826	61%	95 588	41%

additional 10% of all exclusive compounds stem from smaller suppliers (<0.2 million) having over 80% of exclusive species in their individual collections.

The remaining 8% of all exclusive compounds are scattered among 16 suppliers either with very small stocks or with a limited percentage of exclusive species. Therefore, the majority of exclusive compounds can be retrieved from suppliers with the largest libraries and from those focused on offering largely (only) exclusive substances.

Drug-likeness Analysis. The detailed results of applying the Lipinski and Veber rules comprising eight different filters to 5 183 506 compounds from 29 suppliers are given in Figure 2 and Table 2. The inspection of the diagrams shown in Figure 2 reveals that very similar fractions of compounds from all suppliers pass filters accounting for the numbers of hydrogen-bond acceptors and donors, rotatable bonds, and polar surface area and for the Caco2 membrane permeability.

Filters that introduce noticeable differentiation among the suppliers are those taking into account molecular weight, solubility in water, and Clog *P*. For example, more than 90% of compounds from 12 suppliers pass the molecular weight filter. On the other hand, there are six suppliers having 60−80% of their compounds passing the molecular weight filter. Databases that contain the largest number of compounds with molecular weight >500 are Oakwood Chemicals, Fluorochem, Toslab, Albany, and Ufark and the open NCI database. Databases featuring the best passing scores (>90%), i.e., the minimum number of compounds with molecular weight >500 are Asinex, Asis Chem, ChemBridge, ChemDiv, Enamine, InterBioScreen, IVK Laboratories, KeyOrganics, Peakdale, and Tripos. In the case of the Clog *P* filter, the passing percentage of different suppliers' libraries range from 62 to 92%. Databases that show the highest level of violations on the basis of the distribution of Clog *P* are Albany, ChemStar, Otava Chemicals, SALOR, Pharmeks, Princeton Biomolecular Research, and Ufark. The latter libraries have the Clog *P* filter passing percentage below 70%. The catalogues with the best Clog *P* passing scores (>90%) are those of Enamine, Fluorochem, and Oakwood Chemicals. Some 68−98% of the compounds of different suppliers pass the solubility (log *S*) filter. The database of ChemStar shows the highest failure rate on the basis of the calculated log *S*. The databases of Fluorochem and Oakwood Chemicals and the open NCI reveal the best results in passing the solubility filter. The results of applying the filter that removes toxic and chemically reactive species from the databases are given in Table 2. The latter filter has the most dramatic action on the databases of some suppliers. For example, there are five suppliers whose databases pass this filter below the 60% level. At the same time, only eight suppliers have over 80% of their compounds passing this filter. The action of sets of filters, grouped by the rules of Lipinski and Veber, as well as percentages of compounds that passed all drug-like filters are also given in Table 2. Compared to the 2004 analysis of Baurin et al.,[13] a higher percentage of compounds satisfy the Lipinski rules and a lower percentage of compounds pass the Veber filter. The percentage of the compounds that pass all drug-like filters has increased since the 2004 report (39

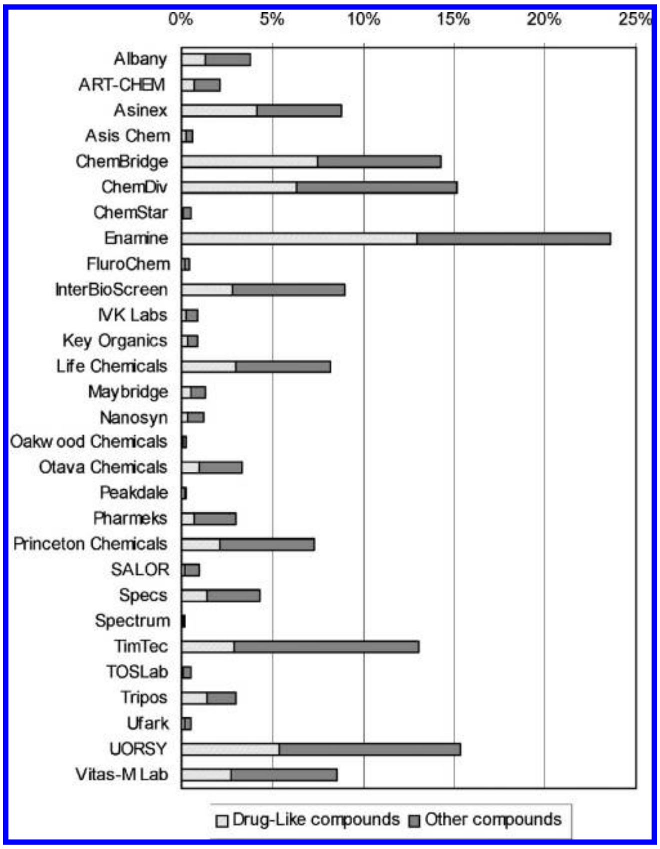

Figure 3. Contributions of individual suppliers to the combined chemical database.

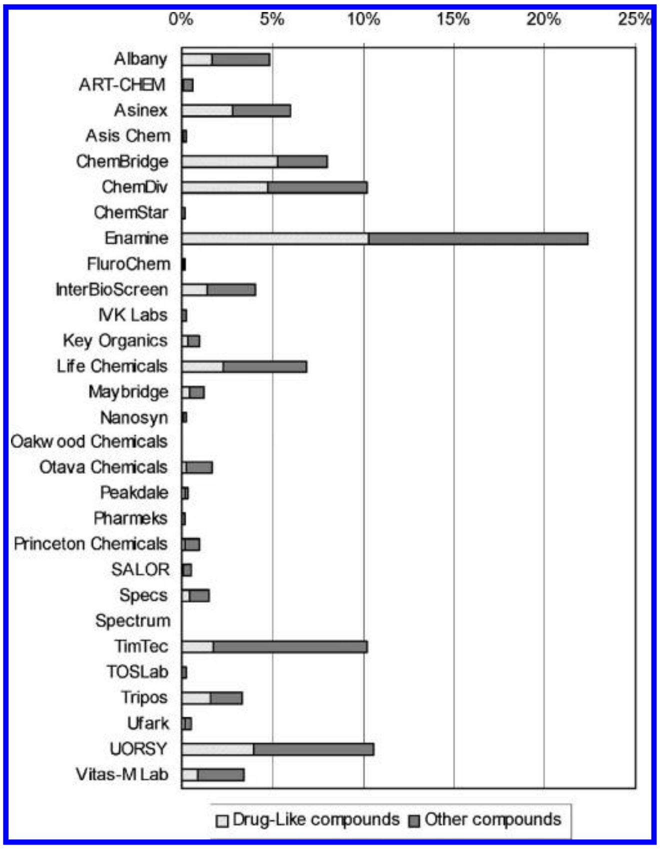

Figure 4. Exclusivity diagram showing both general and drug-like compound availabilities from the 29 suppliers.

vs 37%). This indicates that, in general, suppliers have been paying more attention to the drug-likeness of their compounds.

Databases of ChemBridge and Enamine have the best overall drug-like filter passing scores (>50%), whereas libraries of ChemStar and Princeton Biomolecular Research have the highest number of violations (<20% drug-like). Figure 3 presents a diagram showing the contributions of individual suppliers to the whole chemical database analyzed in this work. It is seen that the 10 largest suppliers contribute more than 90% of compounds to the combined database. Moreover, the 10 largest suppliers provide also more than 90% of the drug-like compounds.

Table 3 and Figure 4 show the results of applying the filters to the compounds that are exclusive to a given supplier. In this case, only 8 of the 10 largest suppliers are in the 'top 10' with regard to the exclusivity. Albany and Tripos with relatively small libraries (<0.2 million) containing 99 and 88% of exclusive species, respectively, showed remarkable contribution to the total amount of exclusive drug-like compounds. Nevertheless, the 10 largest suppliers hold more than 90% of the drug-like exclusive compounds.

Lead-likeness Analysis. Real lead structures are usually synthetically modified in a systematic way before they may turn to biologically active substances for clinical tests. The synthetic modification of the lead implies certain variations of its physical properties, e.g., increase in molecular weight and lipophilicity to values on the order of drug-likeness parameters. This is why, compared with drug-like molecules, the identification of lead-like species requires applying of more rigorous property cutoff values in the database filtration process (see Experimental Section). The inspection of results of the lead-likeness analysis given in Table 3 reveals that the population of the lead-like structures in the whole database is expectedly smaller than that of the drug-like ones (32 vs 39%). Additionally, only one-half (16%) of the lead-like species are exclusive to some supplier. This is in stark contrast to the drug-like collection of which the majority belongs to the exclusive set. Another interesting trend is relative availabilities of the lead- and drug-like compounds from particular suppliers. As seen from Tables 3 and 4, the amount of the lead-like species in the stocks of a majority of suppliers is smaller, by nearly a factor of 2, than their stock amount of the drug-like compounds. Only a few suppliers, e.g., Fluorochem, Oakwood Chemicals, and Ufark have comparable numbers of the lead- and drug-like species in their stocks. Other trends regarding availabilities of the lead-like structures are similar to those of the drug-like ones discussed above. For example, the 10 largest suppliers hold more than 90% of the lead-like exclusive compounds.

Analysis of Biological Activity and Structural Features. The estimated biological activities of the database compounds are depicted in Figure 5. All suppliers have very similar proportions of potentially biologically active species with respect to a particular class of bioactivity. Thus, a noticeable percentage of compounds with potential antiviral activity and with GPCR and protease binders is typical for all suppliers' stocks. At the same time, all the stocks are lacking prospective ion channel modulators and kinase inhibitors. Figure 6 supplements the information on the exclusively provided bioactive species. The bioactive species are often shared by two or more suppliers. In this case, the situation is similar to that of the lead-like compound availabilities.

Table 3. Properties and Exclusivity of the Combined Database with Regard to the Lead-likeness

		total filtered lead-like		total filtered lead-like, exclusive		
supplier	library size	compounds	%	compounds	%	% of total lead-like exclusives
Albany Molecular Research	196 064	31 512	16%	30 773	16%	3.6%
ART-CHEM	110 873	21 882	20%	3215	3%	0.4%
Asinex	457 842	131 093	29%	66 006	14%	7.7%
Asis Chem	32 749	7420	23%	1876	6%	0.2%
ChemBridge	741 176	250 740	34%	141 438	19%	16.6%
ChemDiv	785 740	188 588	24%	107 468	14%	12.6%
ChemStar	28 946	2850	10%	400	1%	0.0%
Enamine	1 221 957	291 751	24%	223 668	18%	26.3%
FluoroChem	23 498	6435	27%	1472	6%	0.2%
InterBioScreen	466 671	81 515	17%	29 490	6%	3.5%
IVK Laboratories	46 515	8877	19%	1195	3%	0.1%
Key Organics	47 656	12 094	25%	9458	20%	1.1%
Life Chemicals	426 135	76 726	18%	43 187	10%	5.1%
Maybridge	69 862	17 533	25%	11 432	16%	1.3%
Nanosyn	62 597	11 354	18%	1852	3%	0.2%
Oakwood Chemicals	12 621	3305	26%	197	2%	0.0%
Otava Chemicals	173 941	27 086	16%	6325	4%	0.7%
Peakdale	14 576	4373	30%	4291	29%	0.5%
Pharmeks	155 800	18 949	12%	367	0%	0.0%
Princeton Biomolecular Research	380 424	62 285	16%	4289	1%	0.5%
SALOR	48 693	5638	12%	2925	6%	0.3%
Specs	223 630	43 042	19%	9837	4%	1.2%
Spectrum	8497	2051	24%	33	0%	0.0%
TimTec	674 773	86 482	13%	38 256	6%	4.5%
TOSLab	26713	2934	11%	835	3%	0.1%
Tripos	154 604	44 159	29%	38 448	25%	4.5%
Ufark	28 881	5808	20%	4135	14%	0.5%
UORSY	794 997	118 159	15%	53 939	7%	6.3%
Vitas-M Lab	442 971	80 961	18%	15 147	3%	1.8%
unique	5 183 506	1 645 602	32%	851 954	16%	
open NCI database compounds	231458	38613	17%	31745	14%	

The chemical character of the database compounds presented in Figure 7 is illustrated by 25 of the most frequently met cyclic fragments and is sorted by individual suppliers. Noteworthy, ca. 90% of the structures of all suppliers contain a benzene ring. The other fragments are very supplier dependent. This collection of data can serve as a practical source of information for customers looking for specific chemical features to fit a concrete drug discovery project.

Analysis of molecular diversity. Since high structural diversity of the compound libraries is as important as their drug- and lead-likeness, we performed an estimation of molecular diversity of the assembled database in terms of both diversity coefficient and clustering (see Table 4 and Figure 8). The results of clustering include: (i) cluster size (small, medium, and large) and percentage in suppliers' database, (ii) average cluster size that is indicative of dominating cluster size in a compound collection, and (iii) number of singletons highlighting under-represented compound types. In terms of diversity coefficients listed in Table 4, all suppliers' databases reveal high structural diversity. Irrespective of the library size, all databases have very close values to their diversity coefficients. This observation indicates that all suppliers have been generally focused on increasing the structural assortment of their libraries. However, the results of clustering reveal considerable differences among suppliers' collections. A large amount of small clusters in a compound library is a hallmark of its high structural diversity. Nevertheless, certain care in analyzing the clustering data must be taken during the selection process. As seen from Table 4, although the high percentage of small clusters is characteristic for almost all suppliers, the average cluster size varies considerably from 3.7 to 26.7. This observation is explained by the diagram in Figure 8, showing that the percentage of compounds belonging to a certain cluster size is extremely supplier dependent. Therefore, while selecting compounds for screening purposes, it is of importance to monitor percentages of the number of both the clusters and the compounds in a given cluster size. The average cluster-size value serves as a useful indicator in the

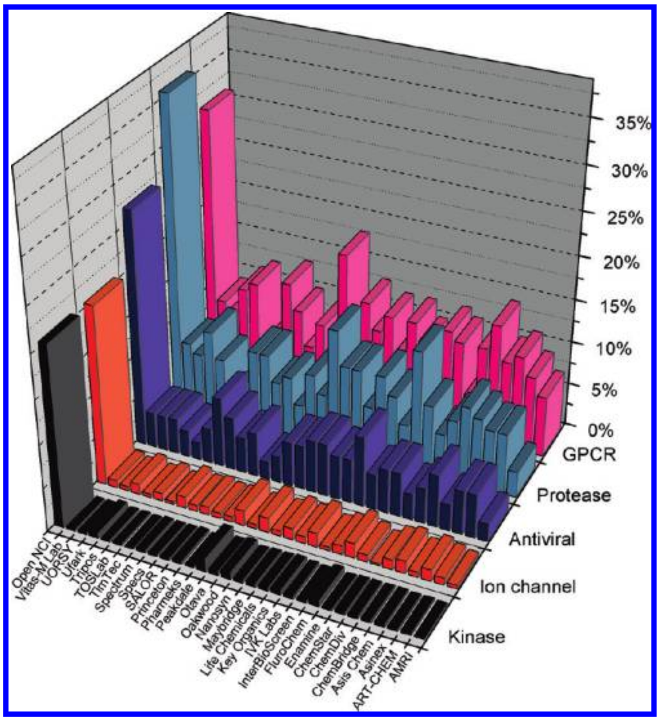

Figure 5. Assessed biological activities of compounds from the assembled database.

Table 4. Diversity Analysis of the Assembled Database

		number of clusters									
supplier	diversity coefficient	small (1−10)	%	medium (11−100)	%	large(>100)	%	total	singleton s	%	average cluster size
AMRI	0.812	6503	69.2%	2521	26.8%	372	4.0%	9396	2140	1.1%	26.7
ART-CHEM	0.801	21 277	91.1%	1995	8.5%	71	0.3%	23 343	10 775	9.7%	8
Asinex	0.809	70 521	89.5%	7892	10.0%	425	0.5%	78 838	35 286	7.7%	9.7
Asis Chem	0.809	4888	89.5%	547	10.0%	29	0.5%	5464	2436	7.4%	10
ChemBridge	0.811	129 317	90.2%	13 629	9.5%	494	0.3%	143 440	63 069	8.5%	8.4
ChemDiv	0.817	88 227	86.5%	12 651	12.4%	1091	1.1%	101 969	42 718	5.4%	12.5
ChemStar	0.785	9587	96.5%	335	3.4%	11	0.1%	9933	5735	19.8%	5.5
Enamine	0.819	224 835	92.2%	18 513	7.6%	515	0.2%	243 863	113 531	9.3%	7.3
FluoroChem	0.805	9015	97.2%	256	2.8%	0	0.0%	9271	5046	21.5%	4.4
InterBioScreen	0.838	62 400	88.9%	7248	10.3%	559	0.8%	70 207	31 413	6.7%	11.2
IVK Laboratories	0.799	11 603	93.5%	792	6.4%	8	0.1%	12 403	5573	12.0%	6
Key Organics	0.818	17 286	96.5%	623	3.5%	1	0.0%	17 910	9783	20.5%	4.7
Life Chemicals	0.794	30 777	82.2%	5967	15.9%	719	1.9%	37 463	13 946	3.3%	17.5
Maybridge	0.829	31 744	98.5%	490	1.5%	0	0.0%	32 234	18 471	26.4%	3.7
Nanosyn	0.811	18 692	95.4%	883	4.5%	19	0.1%	19 594	10 981	17.5%	6
Oakwood Chemicals	0.843	5451	97.8%	123	2.2%	0	0.0%	5574	3284	26.0%	4.1
Otava Chemicals	0.793	24 564	88.8%	2905	10.5%	189	0.7%	27 658	12 664	7.3%	10.8
Peakdale	0.807	3284	92.3%	272	7.6%	1	0.0%	3557	1325	9.1%	5.9
Pharmeks	0.806	29 874	92.0%	2472	7.6%	135	0.4%	32 481	16 499	10.6%	8.7
Princeton Biomolecular Research	0.812	54 800	88.6%	6675	10.8%	354	0.6%	61 829	26 657	7.0%	10.1
SALOR	0.837	16 726	96.7%	553	3.2%	21	0.1%	17 300	10 744	22.1%	5.8
Specs	0.821	56 155	93.9%	3528	5.9%	120	0.2%	59 803	31 763	14.2%	6.8
Spectrum	0.799	2187	94.2%	131	5.6%	3	0.1%	2321	1114	13.1%	6.1
TimTec	0.808	77 468	85.5%	12 462	13.7%	722	0.8%	90 652	37 291	5.5%	11.9
TOSLab	0.798	6148	93.3%	425	6.4%	17	0.3%	6590	3526	13.2%	7.6
Tripos	0.826	31 623	91.5%	2856	8.3%	88	0.3%	34 567	18 553	12.0%	8.5
Ufark	0.805	12 841	97.9%	268	2.0%	3	0.0%	13 112	8034	27.8%	4.1
UORSY	0.803	139 840	91.5%	12 660	8.3%	308	0.2%	152 808	70 096	8.8%	7.5
Vitas-M Lab	0.791	57 004	87.4%	7793	11.9%	430	0.7%	65 227	26 706	6.0%	10.8
joined	0.803	688 066	87.2%	94 135	11.9%	6906	0.9%	789 107	324 905	6.3%	11.8
open NCI	0.897	91 475	97.1%	2762	2.9%	16	0.0%	94 253	55 916	24.2%	4.6

selection process. Setting a concrete cut-off value for the optimum average cluster size depends on the purpose for which a compound library is selected. Compound collections selected for screening purposes usually have the average cluster size value lying within 2−5. The percentage of singletons per database also fluctuates considerably, from 1.1 to 27.8%. Neither a high nor a low percentage of singletons is a danger to a compound library, providing the rest of the library is diverse. Therefore, the number of singletons delivers valuable information only in connection with the rest of clustering data for a given supplier. Notably, the open NCI database has been intended to have the highest diversity in terms of both diversity coefficient and clustering data. It serves, therefore, as a good benchmark for commercial libraries.

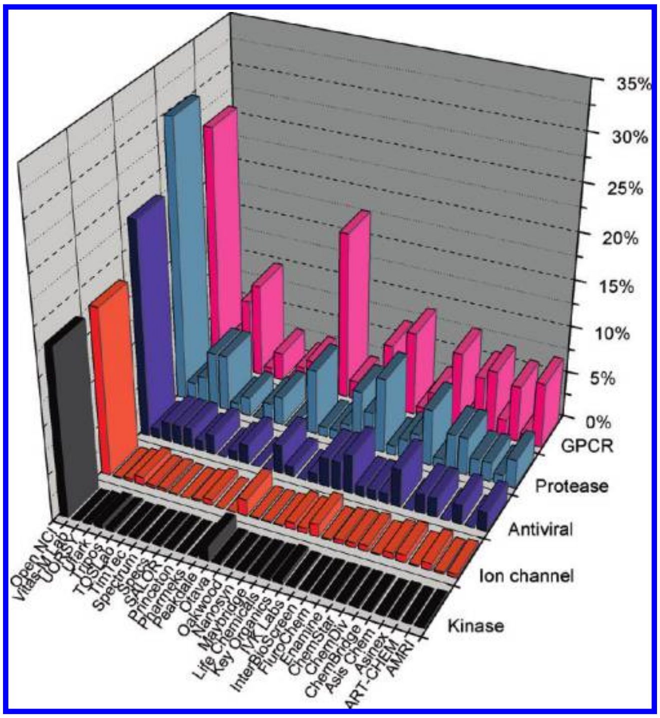

Figure 6. Exclusively offered compounds which were assessed to be biologically active.

CONCLUSIONS

The following conclusions can be drawn on the basis of the presented analysis. The ratio between the number of unique structures and the total number of structures in a joint database has increased in the last five years. This indicates that suppliers have paid certain attention on having more exclusive compounds in their catalogues. Our study has shown that the majority of exclusive compounds can be retrieved from suppliers with the largest databases and from small ones, which are focused on offering only exclusive substances. A detailed view of how the compound libraries pass different drug-like filters reveals that there are four filters that introduce noticeable differentiation among the suppliers' libraries. These filters take into account the molecular weight range, the aqueous solubility, the partition coefficient, and the presence of toxic and reactive groups. The fact that the percentage of the compounds that pass all drug-like filters has shown some increase in the recent years indicates that suppliers have paid specific attention to the drug-likeness of their compounds. On the contrary, the percentage of the lead-like structures has showed some decrease since 2006. Another observation indicating that all suppliers have focused on improving the quality of their libraries by adding valuable

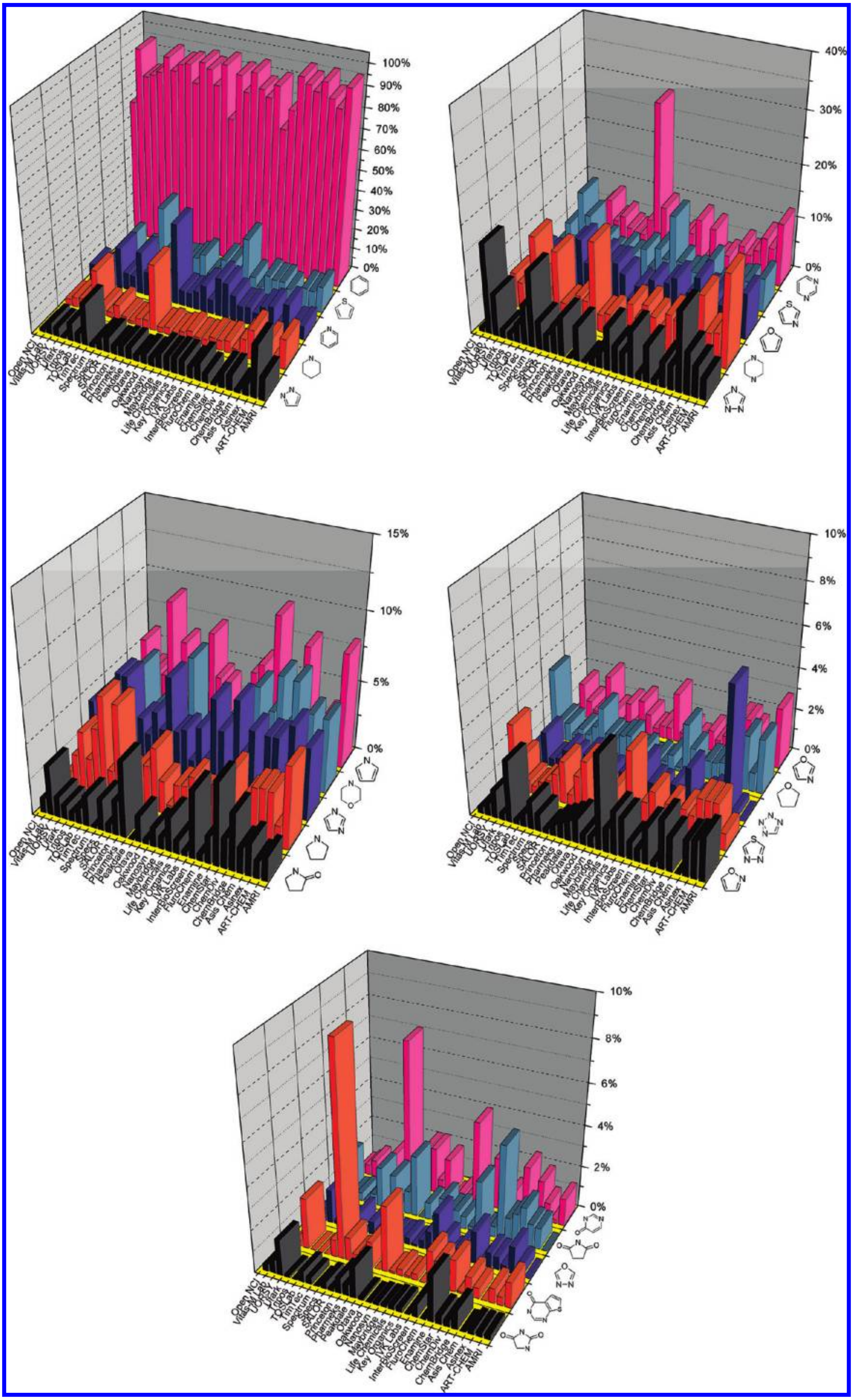

Figure 7. Structural character of suppliers' compounds.

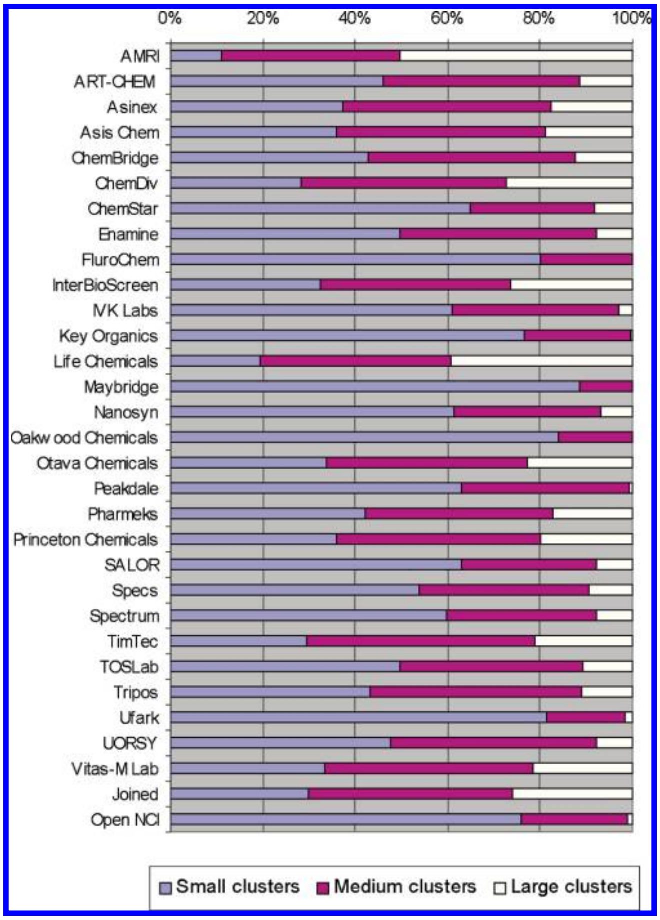

Figure 8. Percentage of compounds in corresponding clusters. Cluster classification is following: small clusters include 1−10 compounds; medium clusters comprise 11−100 compounds; and large clusters contain more than 100 representatives.

compounds is the considerably high values of their structural diversity coefficients and the high percentage of small clusters in their libraries. At the same time, the results of clustering revealed considerable differences among suppliers' collections. The most striking differences were found in the average cluster sizes and the numbers of singletons. Our results show that 10 suppliers with the most sized compound libraries contribute more than 90% of compounds to the 29 supplier database. The 10 largest suppliers hold also more than 90% of the total filtered drug- and lead-like compounds. Finally, the contemporary and comprehensive source of reference on potential biological activities as well as structural character of the database compounds presented in this work can be useful for chemists and biologists responsible for selection of compound libraries for drug discovery projects.

Supporting Information Available: Tables listing detailed data represented in the form of diagrams. This information is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Hüser, J.; Lohrman, E.; Kalthof, B.; Burkhardt, N.; Brüggemeier, U. Bechem, M. High-throughput Screening for Targeted Lead Discovery. In *High-Throughput Screening in Drug Discovery*; Hüser, J., Ed.; Wiley-VCH: Weinheim, Germany, 2006; pp 15−36.

(2) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55–68.

(3) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251–264.

(4) Feher, M.; Schmidt, J. M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.

(5) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeny, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(6) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, 2615–2623.

(7) Livingstone, D. J.; Ford, M. G.; Huuskonen, J. J.; Salt, D. W. Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 741–752.

(8) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077–1084.

(9) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.

(10) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208–1217.

(11) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439–445.

(12) Butina, D.; Gola, J. M. Modeling Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837–841.

(13) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greany, P.; Morley, D.; Hubbard, R. E. Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643–651.

(14) Siroisa, S.; Hatzakisa, G.; Weic, D.; Qishi Duc, Q.; Chou, K.-C. Assessment of Chemical Libraries for Their Druggability. *Comput. Biol. Chem.* **2005**, *29*, 55–67.

(15) Verheij, H. J. Leadlikeness and Structural Diversity of Synthetic Screening Libraries. *Mol. Diversity* **2006**, *10*, 377–388.

(16) *MySQL*, version 5.0.45-community-nt; MySQL AB: Cupertino, CA, 2008; http://www.mysql.com. Accessed July 12, 2009.

(17) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(18) *JChem*, version 5.1.4; ChemAxon Kft: Budapest, Hungary, 2008; http://www.jchem.com. Accessed December 28, 2009.

(19) *LigPrep*, version 1.0.010, and *QikProp*, version 2.1.008; Schrödinger LLC: San Diego, CA; 2008; http://www.schrodinger.com. Accessed July 16, 2009.

(20) Stepanchikova, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Prediction of biological activity spectra for substances: Evaluation on the diverse set of drugs-like structures. *Curr. Med. Chem.* **2003**, *10*, 225–233.

(21) Trepalin, S. V.; Yarkov, A. V. CheD: Chemical Database Compilation Tool, Internet Server, and Client for SQL Servers. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 100–107.

(22) *ISIS/Base*, version 2.4; Symyx Technologies Inc.: Sunnyvale, CA, 2008; http://www.symyx.com. Accessed July 16, 2009.

(23) *MolCart Compounds*, version 200810; Molsoft LLC: La Jolla, CA, 2008; http://www.molsoft.com. Accessed December 28, 2009.

(24) Hann, M. M.; Oprea, T. I. Pursuing the leadlineness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255–263.

(25) Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 667–670.

(26) Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quant. Struct.-Act. Relat.* **1996**, *15*, 285–289.

CI900464S