J|A|C|S
COMMUNICATIONS

Published on Web 06/08/2009

970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13

Lorenz C. Blum and Jean-Louis Reymond*

Department of Chemistry and Biochemistry, University of Berne, Freiestrasse 3, CH-3012 Berne, Switzerland

Received March 24, 2009; E-mail: jean-louis.reymond@ioc.unibe.ch

One of the most important chemical issues in drug discovery is innovation, in particular at the level of small organic fragments that can provide new lead structures.[1] The search for novel molecules can be assisted by in silico methods such as enumeration of chemical space,[2,3] breeding of molecules by genetic algorithms,[4] and analysis of molecular scaffolds.[5] We recently proposed an exhaustive enumeration approach for small organic molecules by assembling the chemical universe database GDB-11,[6] which describes the 26.4 million structures containing up to 11 atoms of C, N, O, and F that satisfy simple chemical stability and synthetic feasibility rules. We now report GDB-13, which enumerates in a similar manner small organic molecules containing up to 13 atoms of C, N, O, S, and Cl. With 977 468 314 structures, GDB-13 is the largest freely available small molecule database to date.

Table 1. Structure Generation Statistics for GDB-13

nodes[a]	graphs[b]	GDB[c]	Cl/S[d]	CPU time (h)[e]
1	1	1	0	0.00
2	1	3	0	0.00
3	2	12	0	0.00
4	4	43	0	0.00
5	8	155	3	0.01
6	20	934	19	0.02
7	57	5 726	315	0.05
8	194	37 151	2 438	0.33
9	706	255 542	17 056	2.68
10	2 831	1 784 626	130 465	25.26
11	12 011	12 961 686	938 704	223.49
12	53 789	99 821 343	7 240 108	3 023.79
13	250 268	795 244 451	59 027 533	36 606.45
Total	**319 892**	**910 111 673**	**67 356 641**	**39 882.08**

[a] Number of graph nodes considered. [b] Number of graphs corresponding to saturated hydrocarbons passing topological and ring-strain criteria. [c] Molecules obtained from the graphs by combinatorial enumeration of unsaturations and heteroatoms and satisfying chemical stability and synthetic feasibility criteria. [d] Molecules with a selection of Cl/S-containing functional groups (see the text and Supporting Information for details). [e] The database was computed in parallel on a 500-node cluster (see the Supporting Information for details).

The assembly of our previously reported GDB-11 started with a collection of graphs[7] considered as hydrocarbons, from which chemically relevant cases were selected by topological and ring-strain criteria and expanded to produce more molecules by introducing unsaturations and heteroatoms following valency rules.[6] The limiting factor in computing GDB-11 was the elimination from this initial list of 98.4% of unstable and/or chemically impossible molecules using functional-group filters. Because most of the rejected molecules contained multiple heteroatoms, we reasoned that it might be possible to accelerate the database computation using a very fast "element-ratio" filter. Analysis of databases of known compounds suggested cutoff values of $(N + O)/C < 1.0$, $N/C < 0.571$, and $O/C < 0.666$ (see the Supporting Information). We also eliminated fluorine because it was rarely found and never considered in our group for synthesis in virtual-screening guided drug discovery applications of GDB-11.[8] Together with the optimization of graph selection by replacing the computationally slow MM2 minimization[9] with a simple geometry-based estimation of strained polycyclic ring systems (see the Supporting Information) and some general code improvement, the assembly time for GDB-11 was thus reduced 6.4-fold, from 1600 to 250 CPU h.

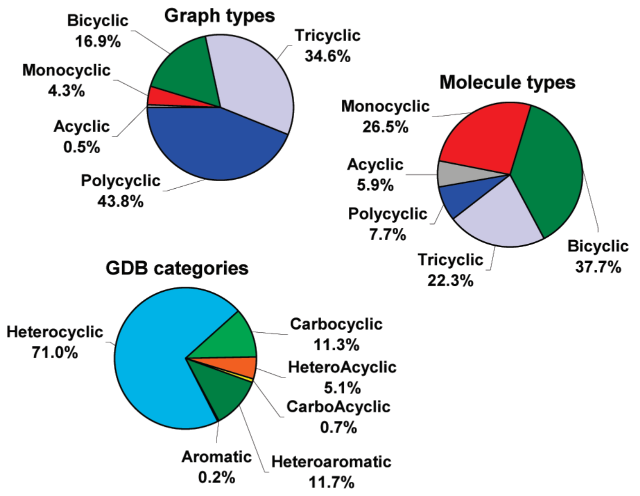

Figure 1. Composition of GDB-13. Category priority: heteroaromatic > aromatic > heterocyclic > carbocyclic > heteroacyclic (interrupted carbon chain) > carboacyclic (continuous carbon chain).

With these improvements, the algorithm was sufficiently fast to compute the database up to 13 atoms, which produced 910 million molecules in 40 000 CPU h (Table 1). In addition, we also produced a chlorine/sulfur set of 67.3 million compounds that enumerates all molecules up to 13 atoms with sulfur atoms appearing in aromatic heterocycles (e.g., thiophenes), sulfones, sulfonamides, and thioureas and chlorine atoms as aromatic substituents. The Cl/S set is of interest for virtual screening because of the distinct molecular shapes and functional groups that are possible with these larger atoms.

The molecular diversity of GDB-13 is well-illustrated by the available molecular types (Figure 1). While polycyclic topologies dominate the graphs, the molecular enumeration results in a majority of monocyclic, bicyclic, and tricyclic molecules, most of which are heterocyclic; 54% of GDB-13 molecules have at least one three- or four-membered ring. The distribution of descriptor values shows that essentially all the molecules are druglike according to Lipinski[10a] (100%) or Vieth[10b] (99.5%). Many of them are also leadlike[10c] (98.9%) or fragmentlike[10d] (45.1%) (Figure 2).[11]

The size of GDB-13 is a consequence of the systematic combinatorial enumeration. For example, between 0.2 and 18 million compounds share the structural formula of typical marketed drugs present in GDB-13, some of which are structurally very

10.1021/ja902302h CCC: $40.75

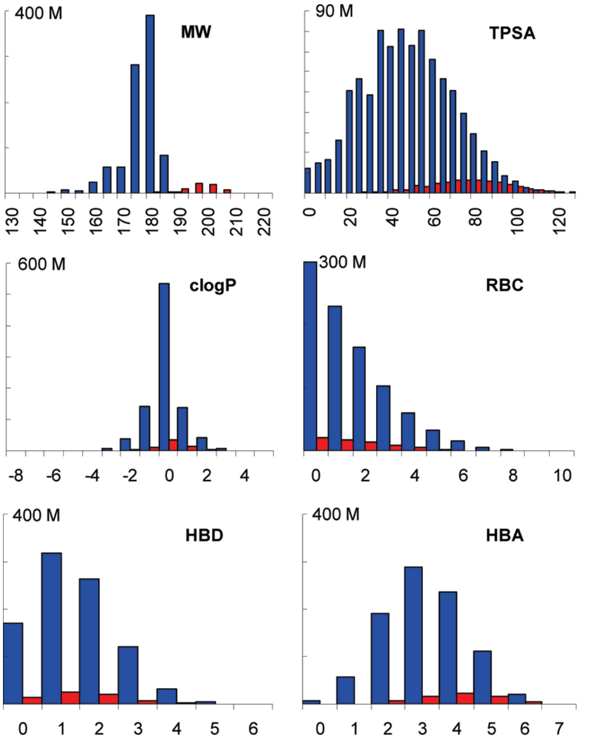

Figure 2. Distribution of C/N/O molecules (blue bars) and the Cl/S set (red bars) in GDB-13 according to property values. MW = molecular weight in Da. TPSA = topological polar surface area in Å^2.[12] clogP = calculated water/octanol partition coefficient. RBC = rotatable bond count. HBD/A = hydrogen-bond donor/acceptor atom count.

Table 2. Structural Isomers of Marketed Drugs Found in GDB-13

			T_{SF}[c]	
name[a]	formula	same formula[b]	avg	>0.7
aspirin	$C_9H_8O_4$	804 153	0.23	178
benzocaine	$C_9H_{11}NO_2$	1 846 579	0.24	74
L-tyrosine	$C_9H_{11}NO_3$	9 276 529	0.46	24 952
levetiracetam	$C_8H_{14}N_2O_2$	2 154 955	0.28	35
memantine	$C_{12}H_{21}N$	2 872 586	0.31	10 912
menadione	$C_{11}H_8O_2$	233 715	0.44	112 186
metaraminol	$C_9H_{13}NO_2$	2 920 516	0.26	30
mexiletine	$C_{11}H_{17}NO$	18 371 393	0.25	119
propofol	$C_{12}H_{18}O$	5 263 227	0.25	240
rasagiline	$C_{12}H_{13}N$	1 323 525	0.13	411
rimantadine	$C_{12}H_{21}N$	2 872 586	0.26	173

[a] Common drug names as found in the DrugBank database.[13] [b] Number of GDB-13 molecules sharing the same structural formula. [c] Tanimoto similarity[14] compared to the parent drug; "avg" = average value across all compounds sharing the structural formula, ">0.7" = number of these compounds with a Tanimoto value greater than 0.7.

similar to the parent compounds as estimated by their Tanimoto coefficients of structural fingerprints (Table 2).[14]

On the other hand, GDB-13 leaves out a large fraction of chemical space because of the choices made to accelerate computation. Thus, of the 619 675 structures containing up to 13 atoms that are found in PubChem,[15] ACX,[16] and the NCI Open Database,[17] 66.2% do not appear in GDB-13, either because they contain nonenumerated elements [e.g., F, Br, I, P, Si, metals (24.7%)] and functional groups [e.g., chlorine on nonaromatic carbons, mercaptans, sulfoxides, hemiacetals, enamines, allenes (35.9%)] or because their heteroatom-to-carbon ratio is too high [e.g., mannitol (5.3%)] or the parent graph was not considered (0.3%).

Despite these limitations, GDB-13 is to our knowledge the largest publicly available database of virtual molecules ever reported. It contains a wealth of yet unknown structures to be explored and synthesized and should provide a rich source of inspiration for design and synthesis in the search for new bioactive fragments not present in databases of already existing compounds, such as ZINC,[18] ACX, and PubChem. The database is available free of charge at http://www.gdb.unibe.ch.

Acknowledgment. This work was financially supported by the University of Berne and the Swiss National Science Foundation. The authors gratefully acknowledge Chemaxon Ltd. for the donation of the academic license to the JChem package.

Supporting Information Available: Details on database generation, statistical analysis, and samples of the high-similarity structures mentioned in Table 2. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. *Nat. Rev. Drug Discovery* **2003**, *2*, 369.

(2) (a) Lederberg, W. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *53*, 134. (b) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. *J. Am. Chem. Soc.* **1975**, *97*, 5755. (c) Benecke, C.; Grund, R.; Hohberger, R.; Kerber, A.; Laue, R.; Wieland, T. *Anal. Chim. Acta* **1995**, *314*, 141.

(3) (a) Bohacek, R. S.; McMartin, C.; Guida, W. C. *Med. Res. Rev.* **1996**, *16*, 3. (b) Ertl, P.; Jelfs, S.; Muhlbacher, J.; Schuffenhauer, A.; Selzer, P. *J. Med. Chem.* **2006**, *49*, 4568.

(4) (a) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. *Perspect. Drug Discovery Des.* **1995**, *3*, 34. (b) Globus, A.; Lawton, J.; Wipke, T. *Nanotechnology* **1999**, *10*, 290. (c) Douguet, D.; Thoreau, E.; Grassy, G. *J. Comput.-Aided. Mol. Des.* **2000**, *14*, 449. (d) Pegg, S. C.-H.; Haresco, J. J.; Kuntz, I. D. *J. Comput.-Aided. Mol. Des.* **2001**, *15*, 911. (e) Brown, N.; McKay, B.; Gasteiger, J. *J. Comput.-Aided. Mol. Des.* **2004**, *18*, 761. (f) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079. (g) Pierce, A. C.; Rao, G.; Bemis, G. W. *J. Med. Chem.* **2004**, *47*, 2768. (h) Lameijer, E.-W.; Kok, J. N.; Bäck, T.; IJzerman, A. P. *J. Chem. Inf. Model.* **2006**, *46*, 545. (i) van Deursen, R.; Reymond, J.-L. *ChemMedChem* **2007**, *2*, 636.

(5) (a) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17272. (b) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. *J. Chem. Inf. Model.* **2007**, *47*, 47. (c) Pollock, S. N.; Coutsias, E. A.; Wester, M. J.; Oprea, T. I. *J. Chem. Inf. Model.* **2008**, *48*, 1304.

(6) (a) Fink, T.; Bruggesser, H.; Reymond, J.-L. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504. (b) Fink, T.; Reymond, J.-L. *J. Chem. Inf. Model.* **2007**, *47*, 342.

(7) McKay, B. D. *Congr. Numerant.* **1981**, *30*, 45.

(8) Nguyen, K. T.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J.-L. *ChemMedChem* **2008**, *3*, 1520.

(9) Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127.

(10) (a) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3. (b) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. *J. Med. Chem.* **2004**, *47*, 224. (c) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743. (d) Congreve, M.; Carr, R. A.; Murray, C. W.; Jhoti, H. *Drug Discovery Today* **2003**, *8*, 876.

(11) (a) Oprea, T. I. *J. Comput.-Aided. Mol. Des.* **2000**, *14*, 251. (b) Oprea, T. I.; Allu, T. K.; Fara, D. C.; Rad, R. F.; Ostopovici, L.; Bologa, C. G. *J. Comput.-Aided. Mol. Des.* **2007**, *21*, 113.

(12) Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714.

(13) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. *Nucleic Acids Res.* **2008**, *36*, D901.

(14) The standard 512-bit structural fingerprint was used: Chemical Hashed Fingerprints, www.chemaxon.com/jchem/doc/user/fingerprint.html (accessed March 11, 2009). For a review of chemical-similarity searching and similarity measures, see: Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983.

(15) National Center for Biotechnology Information. The PubChem Project. http://pubchem.ncbi.nlm.nih.gov (accessed Aug 4, 2008).

(16) *ChemACX Ultra*, version 8.0; CambridgeSoft Corporation: Cambridge, MA, 2005.

(17) National Cancer Institute. NCI Open Database. http://cactus.nci.nih.gov (accessed Aug 4, 2009).

(18) Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Model.* **2005**, *45*, 177.

JA902302H