Why Does High-Throughput Screening Ever Work?

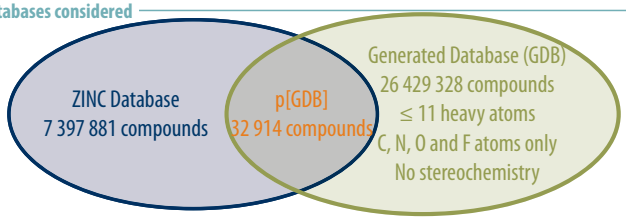
Jérôme Hert and Brian K. Shoichet

University of California San Francisco, 1700 4th St, San Francisco, CA, 94158. Email: hert@cgl.ucsf.edu

INTRODUCTION

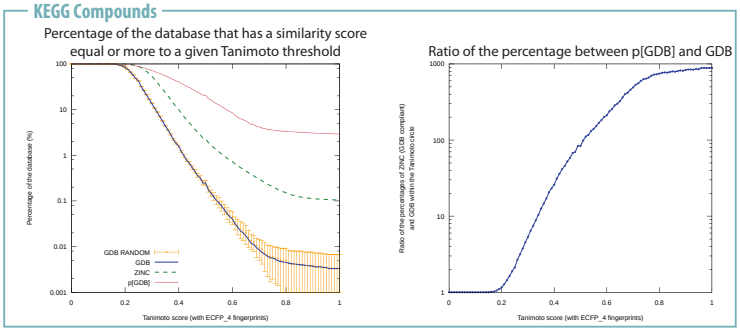
The size of chemical space is believed to range from 10^{60} to 10^{200} molecules. Why should high-throughput screening (HTS), given that it typically considers libraries of 10^6 molecules or less, yield any hit at all? Drug discovery nevertheless heavily relies on HTS and its ability to deliver hits and leads. We believe that screening libraries are biased toward biogenic-like molecules and propose to quantify this assumption. In what follows, chemical space, for molecules of 11 C, N, O or F atoms or less was approximated by the generated database (GDB) of Reymond *et al.*,[1] while screening libraries were approximated by the database of commercially available compounds (ZINC)[2] and its subset that meets the GDB criteria (p[GDB]).

Databases considered

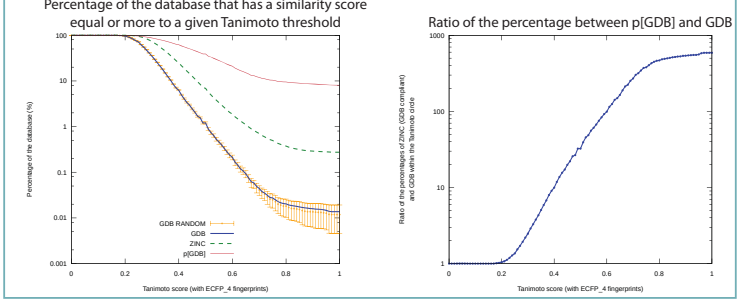


SCREENING LIBRARIES ARE BIASED TOWARD BIOGENIC MOLECULES

KEGG Compounds



Natural products



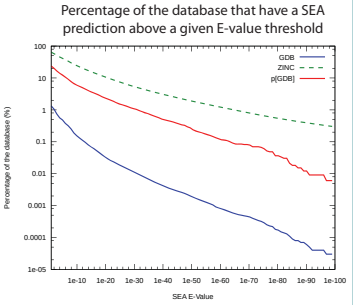
For every compound in the GDB, the p[GDB] and the ZINC database, the maximum similarity to KEGG compounds[3] and natural products[4] was calculated. The left figures show the percentage of the database that had a similarity score equal or better than a given Tanimoto threshold. When KEGG compounds were the reference database, only 5% of the compounds in the GDB database have a similarity equal or higher than 0.34 compared to 57% for the p[GDB] database. At this Tanimoto threshold, the p[GDB] database was hence 10 times as similar to metabolic space as the GDB database. This ratio steadily increased with the Tanimoto threshold to reach almost 3 orders of magnitude between p[GDB] and GDB (right figures). The purchasable compounds were biased toward biogenic molecules: their distribution would be comparable to that of the GDB if they were charting random parts of chemical space (orange lines on the right figures).

REFERENCES

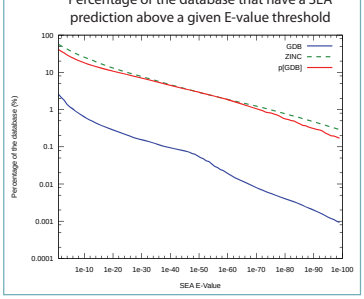
1. T. Fink and J. L. Reymond, J. Chem. Inf. Model. 47 (2), 342 (2007).
2. ZINC is available from http://zinc.docking.org.
3. KEGG is available from http://www.genome.jp/kegg.
4. The Dictionary of Natural Products is available from http://dnp.chemnetbase.com.
5. M J Keiser, B L Roth, B. N. Armbruster et al., Nat. Biotech. 25 (2), 197 (2007).
6. The MDL Drug Data Report database is available from http://mdli.com
7. M. Olah *et al.*, in Chemoinformatics in Drug Discovery, edited by T.I. Oprea (Wiley-VCH, New York, 2004), pp. 223.

SCREENING LIBRARIES ARE BIASED TOWARD BIOACTIVE MOLECULES

MDDR Schuffenhauer



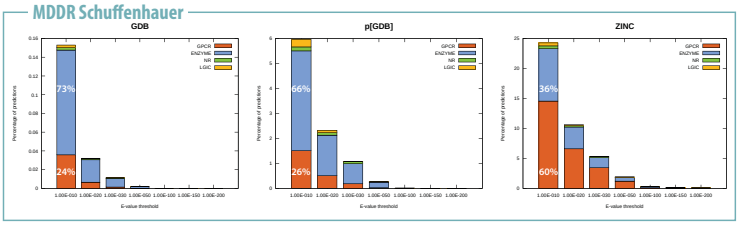
WOMBAT



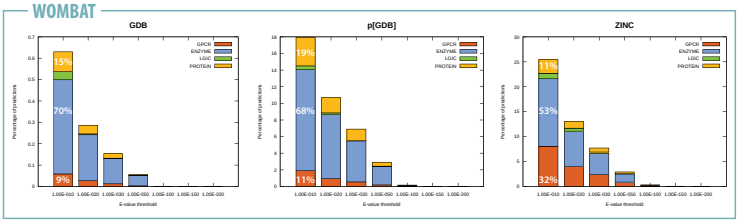
SEA[5] scores were calculated for every compound in the GDB, the p[GDB] and the ZINC database against the ligand sets in the MDDR Schuffenhauer[6] and the WOMBAT databases.[7] 6% of the compounds in the p[GDB] had an E-value of 10^{-10} or better to one of the sets in the MDDR database; in contrast only 0.15% of the compounds in the GDB had a prediction. The ratio of compounds that had a prediction at a given E-value threshold easily reached 2 orders of magnitude between the p[GDB] and GDB databases.

SCREENING LIBRARIES ARE BIASED TOWARD DRUGGABLE TARGETS

MDDR Schuffenhauer



WOMBAT



For the GDB database, 73% and 70% of the targets that were predicted using SEA (with an E-value ≤ 10^{-10}) were enzymes while only 24% and 9% were GPCRs, for the MDDR and the WOMBAT databases, respectively. A similar distribution of target types is observed for the predictions of the p[GDB] compounds. In contrast, 60% and 32% of the targets predicted for the ZINC compounds were GPCRs. Although the p[GDB] database is more similar to biogenic and bioactive molecules than the GBD database it does not bear the same bias toward GPCR like compounds as the ZINC compounds.

WHAT'S MISSING FROM p[GDB] AND GDB?

Minimum Spanning Tree of the MDDR Schuffenhauer Drug Classes