Journal of Computer-Aided Molecular Design, **16:** 301–309, 2002.
KLUWER/ESCOM

An overview of the diversity represented in commercially-available databases

Mary P. Bradley*
Pfizer Global Research and Development, 2800 Plymouth Road, Ann Arbor, MI 48105, USA

Introduction

Commercially available (CA) databases are an important source of starting materials for chemical synthesis as well as compounds for screening. In this overview, 'commercially available database' will be defined as any chemical structure database, either free or for fee, which is generally available to the public, without regard to the availability of the actual compounds contained therein. Analysis of CA databases is generally done in the context of obtaining chemicals from outside sources, either for the purpose of synthetic modification, or screening of the compounds in a biological assay. Searching CA databases for starting materials for synthesis generally does not involve a great deal of analysis over and above a structure or substructure search; simple property filters may be used to eliminate undesirable or unsuitable compounds. In this case, product availability and cost are usually the primary considerations. When the objective is to purchase compounds for screening, the goal in a database analysis may be to enhance the overall quality of oneÆs proprietary screening collection with respect to diversity and/or druglikeness thereby increasing the likelihood that a suitable lead compound will result from screening in biological assays. This can be accomplished by comparing a corporate collection to a standard set with the desired properties (i.e. Drugs) and selecting a subset that has the greatest similarity to the standard set for screening. This might increase the probability that a high quality candidate will be discovered during the screening process. Alternatively, one might wish to enrich the corporate database by adding compounds from outside sources that complement the diversity defined by compounds in the proprietary collection (the *chemistry space*), or explore compounds similar to one that has demonstrated biological activity in an assay to develop a structure-activity relationship.

Thus, the CA databases have been submitted to extensive searching and analysis in a variety of research applications. The majority of the diversity comparisons between CA databases did not begin life explicitly as such. Some are publications of novel metrics or methods of diversity analysis that used CA databases as non-proprietary sources of compounds. There is a wealth of information about how various databases compare with respect to different parameters, metrics, and methods. There have also been numerous analyses of individual CA databases that can serve to lend insight into the composition and internal diversity of these databases. This can be especially helpful when applied to standard CA databases of known drugs, the medicinal chemistry knowledge bases, if one seeks to select compounds that have similar characteristics to known pharmaceutical agents.

Any review of the diversity in CA databases would be remiss without consideration of the impact that the concept of 'druglikeness' has played in such analyses. Since the 'Rule of Five' was introduced by Lipinski et al. [1], these property filters (ClogP $<=5.0$, Molecular Weight $<= 500$, Hydrogen Bond Donors < 5, Hydrogen Bond Acceptors < 10) have become an industry standard for compound selection and acquisition. Other groups have since supplemented these empirical filters to include rotatable bond, rigid bond, and ring counts. [2] The impetus for analyzing CA databases is to find new leads from screening; however the drive to select compounds that fit a druglike profile has shifted the emphasis on diversity to the background. Certainly, placing limits on the size and characteristics of molecules to fit empirical guidelines has the potential to influence the scope of the diversity of new compound collections. While increasing the

To whom correspondence should be addressed. E-mail: Mary.Bradley@pfizer.com chemical diversity of proprietary collections remains an important objective when acquiring compounds, it is no longer the primary consideration. It should be noted as well that druglikeness is itself a property that can be defined and computed on an individual molecule basis [3]. By comparison, diversity is not a molecular property, but a measure of how properties differ between sets. [4] For the purpose of this review, diversity will be defined as the degree of similarity (or difference) between two collections of molecules as a metric relative to some calculated descriptor set.

An overview of the databases

There are four main types of CA database: chemical substance reference databases, fine chemicals databases, databases of biologically active compounds, and screening libraries.

The ultimate sources of compound structures are the chemical substance reference databases, of which the gold standard is the Chemical Abstracts Services (CAS) Registry Database [5] with over 35 million substance records. The CAS database contains all chemical substances that have appeared in the scientific literature from 1957 to the present, with some substances (such as fluorine- and silicon-containing compounds) dating to the early 1900s. [5] CAS, along with Beilstein [6] and SPRESI [7], are general chemical reference databases, and lack information about the availability from commercial sources found in many other databases. In addition, these databases are generally not in a format that is amenable to direct chemical structure analysis, such as pharmacophore searching or property filtering. Any discussion of the diversity of CA databases would indeed be moot when applied to these reference databases, as they contain arguably *all* of the compound structures in the public domain, and therefore define the boundaries of 'real' chemistry space.

The fine chemicals databases include (among others) Available Chemicals Directory (ACD) [8], Specs [9], and Maybridge [10], composed primarily of small molecules suitable for use as starting materials for chemical synthesis, but with little or no claim of bioactivity. Although there may be numerous examples of biologically active compounds in these databases, this is not their primary utility. ACD is a compendium of many suppliers that contains extensive information about quantity, price, and supplier, and facilitates the ordering of compounds from several different sources. Some suppliers may offer on-line internet structure searching with real-time price and quantity information to facilitate compound ordering directly from the supplier. These databases are primarily used by chemists to select and/or place orders for compounds. However, because they are in a format that facilitates chemical structure analysis, and the compounds are readily available, they are a preferred source for compound acquisition, as well as a convenient source of non-proprietary chemical structures.

The medicinal chemistry knowledge bases, i.e. databases such as Modern Drug Data Report (MDDR) and Comprehensive Medicinal Chemistry (CMC) [8], World Drug Index (WDI) [11], the Cambridge Structural Database (CSD) [12], and the National Cancer Institute (NCI) open database [13], contain compounds with known or putative biological activity. These databases are used primarily for reference purposes, though the compounds may also be commercially available. These knowledge bases differ from the reference databases in that they are specifically focused on compounds with claims of biological activity, and are in a format that facilitates chemical structure analysis. They may also contain information about the biological activity and therapeutic targets for the compounds, which could be of interest when using these databases as reference standards. These databases represent the publicly available 'universe' of bioactive compounds, and though the compounds contained are not available (in most cases) from the database supplier, these databases are a primary research tool in the pharmaceutical sciences.

The screening libraries focus on providing compound collections designed specifically for screening. An even further distinction can be made among libraries designed for 'lead finding' and those designed for targeted screening. Targeted screening collections, such as the so-called 'Killer Plates' and Gen-Plus (available from Miscosource) [14] and several platform-specific LIGAND-SET^{TM} collections offered by Sigma-Aldrich [15] comprise molecules whose biological activity and modes of action are well documented. These libraries might well be grouped with the medicinal chemistry knowledge bases; however, they have been specifically compiled and sold for screening, perhaps to validate new biological assays, or enhance SAR. The structures and biological activities of the compounds contained in these targeted screening libraries are for the most part contained in the medicinal chemistry knowledge bases. Lead finding libraries generally contain compounds for which no previous claims of biological activity are made. These collections may be produced using combinatorial synthesis, or compiled as a diverse set from multiple sources. A very few examples of the multitude of available screening libraries include Optiverse (Tripos, Inc.) [16], Compass Array (Arqule) [17], and PHARMACophore (Chembridge) [18]. Generally, one would search these databases for compounds that fit some pre-defined physicochemical criteria to enrich the quality of a proprietary screening collection. The list of suppliers of such screening libraries has increased tremendously over the past few years, and the number of compounds available from these sources today is well over 1 million. [19]. Some companies are beginning to collate these collections in a format this is more amenable to search and analysis. MDL has made available the ACD Screening Compounds collection (ACD-SC) that boasts over 1.8 million unique compound entries from 29 suppliers. ChemNavigator [20] offers an Internet portal for searching over 2 million chemical structures from 40 suppliers. An exhaustive comparison between the CA screening collections has not been made available in the literature, and while it is beyond the scope of this review, it certainly merits further consideration.

How is diversity measured?

Diversity is a relative characteristic, like beauty, that while not totally subjective is highly context-dependent. There are numerous parameters that can be used to describe molecules, and many metrics that can be applied to measure the differences in parameters between molecules. More importantly, there are no hard and fast rules that apply to the use of these metrics or to the interpretation of the results. For example, say one has proposed to make a set of compounds that are found to be diverse from known drug standards. One conclusion that might be drawn is that the compounds are undesirable because they are not similar to a known set of standards with desirable characteristics. On the other hand, the proposed compounds may be equally interesting *because* they are different from known standards, and therefore more likely to represent a novel structural class, or to show activity against a novel target. The interpretation and application of the results of the diversity analysis depends on other contextual influences that are not always quantifiable. Nevertheless, the chemistry space that comprises the universe of all possible molecules is vast, and some regions of this space are more likely to be populated by drugs than others. Defining the characteristics of this region can begin with examination of the compounds that are known to inhabit it already.

When evaluating the diversity of compounds, it is essential to have a frame of reference or standard. Because the Holy Grail of druglikeness looms over each analysis, there must be a reference standard for this property. CMC is the de facto standard for druglike molecules because it entirely comprises marketed pharmaceutical agents. The tacitly unanimous choice for the non-druglike standard appears to be ACD. The choice of ACD as the non-druglike standard is not surprising, as one criterion for druglikeness is the absence of reactive moieties, and ACD is the primary source for synthetic building blocks, which by definition should contain reactive functional groups. However ACD also contains many compounds with known biological activity, and several groups have shown high degrees of similarity between ACD and drug knowledge bases. Whether this is an artifact caused by the use of ACD as the primary source for the synthetic components of drug molecules or of the diversity metrics used has not been explored fully. That ACD has been so widely used in database comparisons is not surprising, as it contains compounds that are readily available for purchase, and all are in the public domain.

The Pesticide Manual [32] is a yearly compendium of marketed pesticides (herbicides, fungicides, insecticides, etc.). This has been used in the past by the author as an alternative standard for non-druglikeness, primarily because nearly all of the compounds in the list are meant to cause fatality of the primary organism. While these compounds have the druglike characteristics of bioavailability and specific mode of action, their human toxicity in most cases qualifies them as non-druglike. It is interesting to note that fewer compounds in the Pesticide Manual than CMC fail the Lipinski rules for druglikeness (unpublished results).

Measures of diversity begin with molecular descriptors. The choice of descriptors can vary widely, and includes calculated properties (ClogP, molecular weight, free energy of solvation, BCUT descriptors), structural motifs (substructure keys, ring-systems), topological indices (atom pairs) [21], MolconnZ, molecular graph descriptors), and pharmacophores. It has been shown by Brown and Martin [22] that 2D substructure keys that are less expensive to use computationally, and can be as powerful as molecular properties and pharmacophores in estimating biological activity. The use of calculated properties for diversity analysis rather than structural descriptors will often give a different picture of diversity, even though the properties are a function of the structure.

Clustering and cell-based methods for quantifying diversity are most commonly used in conjunction with Tanimoto coefficient and Euclidean distance metrics. During the process of drug discovery, many structural analogs of an active or lead compound are synthesized. In the coursed of a diversity analysis, this is evidenced by a non-uniform distribution or clumping of the compounds with similar structures combined with areas of sparse compound habitation. Cell based methods facilitate the selection of compounds that fill these 'holes' in an existing collection and are also useful for determining the overlap between databases. Hierarchical clustering is superior to nearest neighbor methods but more computationally expensive and time consuming, especially for comparing large databases. Examining the distributions of properties is a useful method for evaluating the diversity of compounds *within* a particular database. Alternatively, a pairwise comparison of all compounds in the database can be performed to quantify the diversity of the compounds in the database.

The use of different parameters and metrics can often result in different interpretations of diversity, even between the same compound sets. For example, if one is primarily interested in the differences in structural composition of molecular databases, it makes sense to use descriptors that capture this information, such as molecular structure fingerprints or topological indices. However, as physicochemical parameters are a function of the molecular structure, the diversity of these physicochemical properties may also be captured in such an analysis. The converse of this is not necessarily true, however. Examining the diversity of physicochemical properties may give insight into only the gross features of the compound structures, though this method has less utility for directing the synthesis of new compounds as it does not offer a quantitative assessment of which structural features of a molecule are important to achieve diversity. It can sometimes be challenging to describe calculated properties as structural modifications.

What is the overlap between databases?

As previously noted, many comparisons of CA databases were not meant as scientific explorations of these chemistry spaces per se. However CA databases, the medicinal chemistry knowledge bases in particular, are convenient public domain collections that are frequently used as stand-ins for proprietary compound collections for the purpose of publication. Table 1 gives an overview of some diversity analyses that compare CA databases. Direct comparison of the results of the different analyses is difficult because the databases used in each analysis were not identical, and each analysis took a somewhat different approach in terms of parameters and metrics used. Interpretation of the results can also lead to confusion if not done carefully. If a comparison is made between a large database (e.g. ACD) and a much smaller one (e.g. CMC), then the *percentage* of the smaller database that overlaps with the larger will be more significant. For example, if 2000 compounds in CMC are identical to compounds in ACD, the subset of 2000 represents a much greater percentage of the CMC database than the ACD. This is perhaps a measure more of the *efficiency* of diversity coverage. Smaller, more structurally diverse sets would be expected to have a somewhat higher efficiency than larger sets with many clusters of structurally similar compounds. An example of this is illustrated in Figure 1. Databases B (100,000 compounds) and C (6,000 compounds) are compared with database A. Although fewer compounds in database C *actually* overlap with database A than those in database B, the percentage of compounds in C that overlap (83%) is greater than in B (20%). It is perhaps more consistent to identify the portion of database A that is covered by the respective databases, or to apply some normalization to the analysis.

Voight and coworkers [23] published a comprehensive comparison of the NCI [13] open database with seven other CA databases (five screening libraries and two knowledge bases). Using CACTVS descriptors, fingerprints that contain bits for binned element counts, fragments, and rings, and Tanimoto score to assess the similarity, they found that WDI and NCE each contain only 13 percent of compounds that are *identical* to compounds in ACD, however over 80 percent of compounds in each database showed greater than 0.80 Tanimoto *similarity* to ACD compounds. It should be noted that the similarity of ACD to other CA databases was somewhat lower. Overall they found that CSD contained the most diverse compounds compared to any other database. NCI showed the largest similarity overlap with CSD (72 percent), while Asinex and Maybridge had the lowest similarity (43 and 45 percent, respectively). One of the conclu-

Table 1. Comparison of Commercially Available Databases

Databases	Descriptor	Method	Results	Reference
acd, bio, bra, cmc, cgx, may, mic, mss, opt, rrl, spe	MDL keys	Tanimoto coeff./ J-P Cluster	*Percentage of Compounds Identical to ACD:* opt<mss<bra<spe<cgx<mic<cmc<<rrl=bio<may *Percentage of Compounds Similar to ACD (>=0.75 Tan):* opt<spe=mss=cmc=bra=cgx<mic<bio<may *Percentage of Compounds Similar to CMC (>=0.75 Tan):* mic>bra>mss=cgx=acd>opt=rrl>may=bio	4
nci, cax, may, asx, csd, wdi, sa, acd	CACTVS hash codes	Tanimoto coeff.	*Percentage of Compounds Identical to ACD:* asx<csd<nci=wdi<sa<may=acx *Percentage of Compounds Similar to ACD (>=0.80):* csd<nci=wdi<asx<acx=sa<may *Percentage of Compounds Identical to NCI:* csd<wdi<CCA	23
cmc, mddr, acd, spe	GSOLV, MolconnX	Cell based	*Volume Percent overlap with ACD:* cmc<specs<mddr	25
wdi, acd, csd, nci	ring system/ cluster	DIVPIK	*Diversity of ring clusters:* csd>nci=acd>wdi	24

Database	Key
ACD	acd
CMC	cmc
MDDR	mddr
WDI	wdi
CSD	csd
NCI	nci
ChemACX	cax
Maybridge	may
Asinex	asx
Sigma-Aldrich	sa
Specs	spe
Bionet	bio
Brandon	bra
Comgenex	cgx
Microsource	mic
MSS	mss
Optiverse	opt
Receptor Research	rrl
cax, may, asx, sa	CCA

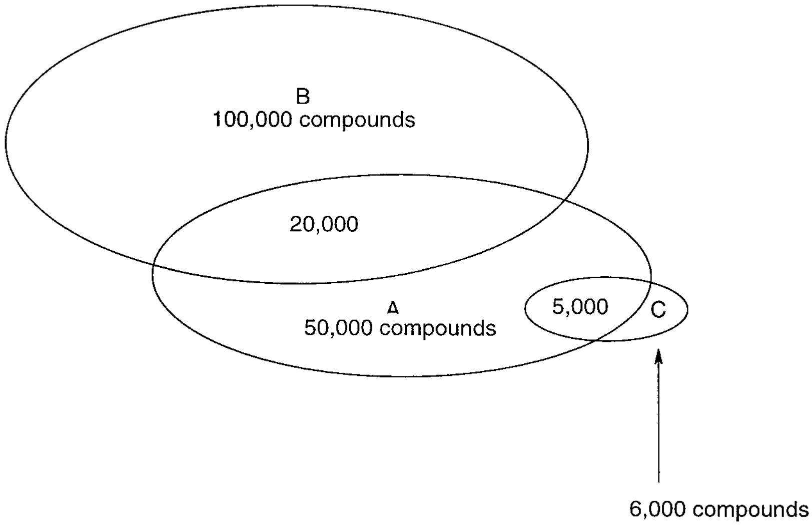

Figure 1. Efficiency of Database Overlap

sions drawn from this study was that database entries with no structure, as well as duplicate entries (such as different salt forms of the same compound), can potentially skew a diversity analysis if they are not removed. The occurrence of duplicate entries was high for the NCI open database because compounds are submitted to the database from many outside sources. Other CA databases, especially ACD, routinely offer different salts of the same parent structure, and likewise contain a significant number of duplicate structures.

Noting that the minority of pharmaceutically active agents are acyclic, Nilakantan et al. used ring content of compounds to identify ring-systems (unique ring types) and ring-clusters (unique combinations of ring-systems) as a method for comparing molecules between databases. They found by this method that CSD contained the richest 'ring' diversity of compounds, with NCI and ACD nearly equal ahead of WDI. They conclude that although there is significant overlap of ring-systems and ring-clusters between databases, each database also contains a substantial proportion of unique ring-systems and ring-clusters. [24]

McGregor and Pallai [4] used MDL substructure keys and Jarvis-Patrick clustering for their analysis of eleven CA databases. Their analysis indicates that there is little overlap of identical compounds between the CA screening libraries and ACD. However they found that all but the Optisim library (54 percent) had a majority of compounds (greater than or equal to 75 percent) with a significant similarity of greater than 0.75 Tanimoto coefficient compared to ACD. When compared to CMC Optisim fared only as well as ACD in terms of the percentage of similar compounds (about 22 percent). Overall, the Optisim library had the most diversity when compared to the other databases, and showed no overlap of identical compounds when compared with other sources.

Cummins et al. [25] employed factor analysis of computed free energy of solvation and 60 topological indices (using Molconn-X) to map compounds into multi-dimensional hypercubes. They then evaluated the volume overlap of each database by counting the number of compounds in each database that occupied each subcube. They further examined the overlap of 'biologically active space' represented by MDDR and CMC to 'commercial compounds' represented by Specs and ACD, and the Wellcome Registry compounds. Their analysis involved the iterative removal of outlier compounds that inflate the descriptor space and effectively compress the remaining compounds into a relatively small volume. Removal of the outliers and re-division of the hyperspace allowed for a more detailed view of the majority of the compounds. Their results indicate that there is significant overlap between all databases studied. 35 percent of the volume occupied by ACD also contains CMC compounds; conversely, 92 percent of the volume occupied by CMC also contains ACD compounds. The immediate conclusion that can be drawn from this is that CMC contains fewer compounds than ACD, and that the diversity in CMC is largely captured by compounds in ACD.

What is the internal diversity? (Or how does druglikeness affect diversity?)

Diversity of CA databases can also be evaluated internally, by examining the extent to which compounds in the database differ from each other. It can be especially useful in the context of drug discovery to evaluate properties of compounds that are already known to have therapeutic activity in humans. When the constraint of druglikeness is imposed upon molecules, and the properties measured to evaluate diversity include druglikeness (or some subset of the spectrum of properties that define it), it may be that internal diversity will be intrinsically reduced. One way to evaluate the internal diversity of databases is by examining the distribution of properties in the database. Perhaps the best known such analysis was done by Lipinski and coworkers [1]. Their observation that compounds in CMC had relatively tight boundaries on four simple molecular properties (Molecular weight, LogP, hydrogen bond donors, and hydrogen bond acceptors) has sparked a revolution in the way that drug design is done. Since these properties are facile and inexpensive to calculate, and intuitive to apply to the design of new compounds, use of the Rule of 5 has been widely embraced by the pharmaceutical research community. The Rule of 5 is not explicitly meant as a means of identifying druglike compounds, however it is quite useful for identifying those compounds for which the 'probability of useful oral activity is low'.

Bemis and Murcko did a further evaluation of CMC to determine the constituent frameworks and side chains that comprise the spectrum of known drugs. [27] Using topological torsions [26] to describe the frameworks, they found that 2506 of these were necessary to describe the 'universe' of known drugs (5120 compounds from CMC). Of these 2506 frameworks, 1908 (76%) occurred only once. 41 frameworks accounted for 24% of the structures studied. They also noted that while the biphenyl framework occurred only 16 times, the compounds in which it was present were distributed among 12 therapeutic classes. The promiscuity of the biphenyl motif was also noted by Hujduk et al. in their study of small molecules binding to proteins using NMR. [28] In a separate study, Bemis and Murcko [29] examined the occurrence of side chains in CMC. They found that out of 5,090 molecules in the study, 4,689 contained side chains (by their definition). Each molecule had an average of 4 side chains; 18,664 side chains were present. Carbonyl (C=O) was the most frequently occurring, with methyl second. They also looked at side chain combinations; C=O, C=O was the most frequently occurring pair. Most interesting to note are the side chain combinations that are not found in CMC. For example, though Cl and Isopropyl were found frequently in CMC (5^{th} and 21^{st} most frequent, respectively), the combination of the two is not observed. Whether this combination in itself is non-druglike, or whether the combination has just not been explored for other reasons is not clear. There is certainly no empirical evidence to suggest that Isopropyl and Cl must not coexist in the same drug molecule!

Oprea [2] examined not only the boundaries on the Rule of Five (and other) parameters, but also the distributions of these properties between MDDR and ACD. He found that a similar percentage of compounds in both databases pass the Rule of Five test, indicating that these rules alone do not discriminate between drugs and non-drugs. However, properties with discrete values (Hydrogen bond donor and acceptor, ring, and bond counts) had a skewed distribution between the two databases, such that ranges could be applied to separate druglike compounds from non-druglike in the two databases. In a separate work, Oprea and co-workers [30] examined the characteristics of those compounds that had been initially classified as 'leads' from pharmaceutical discovery research projects, some of which later went on to be drugs. They found that the initial lead compounds were, on the whole, less complex structurally than drugs, and fell into an even narrower range of parameters than drugs. This finding would imply that when searching for leads from which to make druglike compounds, the size of chemistry space in which to find compounds becomes even smaller.

Muegge and coworkers [31] proposed a simple pharmacophore scheme using four functional motifs combined to form pharmacophores. Their rule-based classification dictates that druglike molecules will have at least two, but fewer than eight of these pharmacophores. Using this rule based filter, nearly 70 percent of compounds in both CMC and MDDR were classified as druglike, while only about 36 percent of compounds in a filtered subset of ACD passed the filter for druglikeness, although reactive compounds and duplicates of CMC and MDDR were removed prior to the analysis. In general, ACD contained a substantially higher number of compounds with 0 or 1 pharmacophore point than the medicinal chemistry knowledge bases, and was lacking in compounds with 3 or greater such points.

Conclusion

Of the three categories of CA databases, the medicinal chemistry knowledge bases, in particular CMC, have been the most extensively scrutinized. Several studies using different metrics have found that CSD contains the greatest structural diversity. Though ACD is most often used as a standard for non-druglike molecules, some studies have found significant similarity between ACD and its druglike standard counterparts, CMC and MDDR. The degree of similarity seems to depend greatly on the choice of parameters and metrics, as well as the degree of initial filtering of the databases prior to analysis. Duplication of structures within the CA databases can be extensive and removal of duplicates prior to any analysis is prudent.

In the quest to determine what characteristics in a molecule enhance its likelihood to become a drug, the medicinal chemistry knowledge bases have been widely scrutinized. From these analyses, it can be seen that though there is diversity in the scaffolds and side chains among drug molecules, there are some preferred structural motifs that appear in drugs with higher frequency. Certainly, side chains that are involved in hydrogen bonding interactions would be expected to found most often in drug molecules, as they form interactions important for binding and selectivity in a drug. Over-interpretation of the diversity content of CMC, and over application of the 'rules' of drugability may lead to self-imposed restriction of the future diversity of new pharmaceutical entities.

It is quite likely that the standards of druglikeness in place today are a function of the historical diversity of biological targets for which they were sought. The concept of *receptor relevant* diversity, which allows the biological target to define the properties of a chemistry sub-space, has been used to help define appropriate metrics for target-specific diversity analysis. [33] Does this concept scale to include *all* biological targets? Have the boundaries of chemical diversity in the CA databases been constrained by the limits in the biological diversity of historical targets, or are the boundaries truly representative of the special characteristics that define a druglike molecule? The human genome project has opened the floodgates of novel biological targets available for research, targets that may have had little exposure in the past. Will this new abundance of biological diversity in targets force a revision of our standards of chemical diversity and druglikeness in small molecules? Some of the property limitations seen in the current CA databases are due to constraints on human bioavailability of small molecules. New drug delivery systems, such as transdermal and inhalation therapies, currently in development may allow the guidelines for bioavailability of drugs to be relaxed. Perhaps large molecular weight molecules, peptides and proteins will be suitable drug candidates in the future if delivery and metabolism are no longer concerns.

The analyses of commercially available databases discussed provide an interesting snapshot of the current content of the history of chemistry and drug discovery. As the sources of commercially available compounds continue to increase, and the variety of therapeutic targets broadens, this historical view will undoubtedly need to be re-evaluated.

References

1. Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., Adv. Drug Disc. Rev., 23 (1997) 3–25.
2. Oprea, T.I., J. Comput.-Aided Mol. Des., 14 (2000), 251–264.
3. Ajay, Walters, P. and Murcko, M.A., J. Med. Chem., 41 (1998) 3314–3324.
4. McGregor, M.J. and Pallai, P.V., J. Chem. Inf. Comput. Sci., 37 (1997) 443–448.
5. Chemical Abstracts Services, 2540 Olentangy River Road, Columbus, OH 43210, http://www.cas.org/EO/regsys.html.
6. Beilstein: http://www.beilstein.com/products/xfire/.
7. InfoChem GmbH, Landsbergerstrasse 408, D-84241 Munich, Germany, http://www.infochem.de/spresi.htm.
8. MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, http://www.mdli.com.
9. Specs: http://www.specs.net.
10. Maybridge: http://www.maybridge.com.
11. Derwent World Drug Index: http://www.derwent.com/world drugindex/index.html.
12. Cambridge Crystallographic Data Center, 12 Union Road, Cambridge, CB2 1EZ, UK: http://www.ccdc.cam.ac.uk/
13. Open NCI Database is available online at http://cactus.nci.nih.gov/ncidb2/.
14. Microsource Discovery Systems, Inc., 21 George Washington Plaza, Gaylordsville, CT 06755. http://msdiscovery.com
15. Sigma-Aldrich Co., http://www.sigmaaldrich.com
16. Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144. http://www.tripos.com.
17. Arqule, 19 Presidential Way, Woburn, MA 01801. http://www.arqule.com.
18. Chembridge Corporation, 16981 Via Tazon, suite G, San Diego, CA 92127. http://www.chembridge.com.
19. Footnote: data compiled from information on compound availability taken from the following companies' web sites: http://www.asinex.com, http://www.chembridge.com, http://www.chemdiv.com, http://www.nanosyn.com, http://www.chemrx.com, http://www.lionbioscience.com, http://www.tripos.com, and http://www.arqule.com.
20. ChemNavigator, 6166 Nancy Ridge Drive, San Diego, CA 92121. http://www.chemnavigator.com.
21. Carhart, R.E., Smith, D.H. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 25 (1984) 64–73.

22. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 37 (1997) 1–9.

23. Voight, J.H., Bienfait, B., Wang, S. and Nicklaus, M.C., J. Chem. Inf. Comput. Sci., 41 (2001) 702–712.

24. Nilakantan, R., Bauman, N. and Haraki, K.S., J. Comput.-Aided Mol. Des., 11 (1997), 447–452.

25. Cummins, D.J., Andrews, C.W., Bentley, J.A. and Cory, M., J. Chem. Inf. Comput. Sci., 36 (1996) 750–763.

26. Nilakantan, R., Bauman, N., Dixon, J.S. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 27 (1987) 82–85.

27. Bemis, G.W. and Murcko, M.a., J. Med. Chem., 39 (1996) 2887–2893.

28. Hajduk, P.J., Bures, M., Preastgaard, J. and Fesik, S., J. Med. Chem., 42 (2000,) 3443–347.

29. Bemis, G.W. and Murcko, M.A., J. Med. Chem., 42 (1999) 5095–5099.

30. Oprea, T.I., Davis, A.M., Teague, S.J. and Leeson, P.D., J. Chem. Inf. Comput. Sci., 41 (2001) 1308–1315.

31. Muegge, I., Heald, S. and Britelli, D., J. Med. Chem., 44 (2001) 1841–1846.

32. The Pesticide Manual is available through The Royal Society of Chemistry. http://www.rsc.org/is/books/pestman.htm.

33. Pearlman, R.S. and Smith, K.M., J. Chem. Inf. Comput. Sci., 39 (1999) 28–35.