

Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics

Christopher Southan*, Péter Várkonyi and Sorel Muresan*
Global Compound Sciences, Computational Chemistry, AstraZeneca R&D, Mölndal

Abstract

The last two years have seen a dramatic expansion in public cheminformatics, as exemplified by the approximate five-fold growth of PubChem from over 50 contributing data sources. Consequently, medicinal chemists who were hitherto limited to commercial databases now also have access to public sources that they can download and/or query directly over the Web. The range of public sources, particularly where they link out to structured bioinformatic and biological data, already offer utilities that have no commercial equivalent. This work reviews compound content comparisons between selected public and commercial databases that capture bioactive content. We focused particularly on those that specify relationships between compounds and their protein targets. Our stringent filtering produced lower unique compound numbers than those reported for individual databases and thereby facilitated standardised comparisons of content. The resultant matrix shows the pairwise comparison of each database and selected subsets. Overall, this showed an unexpected degree of non-overlap, thereby emphasising the complementarity gained from combining public and commercial sources. This conclusion is supported by a Venn-type analysis of GVKBIO, WOMBAT (both commercial) and PubChem (public). These databases show not only overlap but also unique bioactive content in each case because of their different strategies for source selection and data collection.

* Address correspondence to these authors at: Global Compound Sciences, Computational Chemistry, AstraZeneca R&D, Pepparedsleden 1, S-43183 Mölndal, Sweden. Tel +46 (0) 31-706-5891 christopher.southan@astrazeneca.com and Tel +46 (0)31-706-5283 sorel.muresan@astrazeneca.com

1. INTRODUCTION

The disciplines of bioinformatics and cheminformatics have, until recently, developed along different trajectories, even though they are both of crucial importance to medicinal chemistry and its application to the drug discovery process [1]. While there are similarities in underlying informatic concepts one of the key differences is that, while the former developed in the public domain, the latter was largely confined to the commercial sector [2]. This situation changed radically towards the end of 2004 with the arrival of ChEBI and PubChem which independently established formal, searchable links between sequence data hosted in two of the world's leading bioinformatics portals, and structural representations of small molecules that specifically interacted with targets [3]. In the ChEBI case, these were initially a few thousand enzyme substrates and products while PubChem was focused on a larger

starting set of compounds and tumor cell screening data [4,5]. These public databases were not the first to offer searching across links between biological effects and chemical information *per se*. The National Library of Medicine Medical Subject Headings controlled vocabulary indexing of articles for PubMed (MeSH), has linked compound names to papers over many years [6], the Protein Data Bank (PDB) has put much effort into improving the searchability of ligands in protein structure data [7] and the National Cancer Institute's (NCI) anticancer drug-screening database already had advanced chemical data mining utilities [8]. Also pre-dating PubChem are ChemBank and Binding DB that aggregate chemogenomics and *in vitro* compound binding data, respectively [9,10].

However, within less than two years the appearance of ChEBI and PubChem, whilst not necessarily initiating them, has catalyzed no less than four public informatic "revolutions" that will have a very significant impact on medicinal chemistry in both the commercial and academic sectors [11]. These are:

1. The ability to search links between biological effects, protein names, sequence data, and chemical information.
2. Formal embedding of the "missing entity" of chemical structure representation within the global Web of bioinformatic relationships.
3. Deposition, not just of HTS results but also other types of screening data, directly linked to chemical structure information in public repositories[12].
4. Open cheminformatics whereby tools, databases and standards for nomenclature are proliferating in the public domain[13,14].

From its inception, PubChem had the resources and the mandate to incorporate these concepts. It has therefore become a *de facto* global hub with over 50 outlinked sources and additional inlinks from an increasing number of (bio)chemical databases. The sheer variety and proliferation of these defies classification because the represented compounds are connected by different types of informatic relationships that are no longer confined within the traditional boundaries of cheminformatics or bioinformatics[15].

Over the last few years commercial sources have also showed new developments in the range and scope of their products. One of the most valuable of these is the provision of linkage, with varying degrees of specificity, not only between compounds and their biological effects *in vivo* but also with the proteins whose activities they modulate *in vitro*. This is exemplified by two recent databases, the World of Molecular BioAcTivity (WOMBAT) and GVKBIO [16,17]. Both are target-annotated i.e. they include specific links between compounds and sequence identifiers for the proteins against which these are targeted. Historically, medicinal chemists have explored such relationships by keyword searching the literature and finding, for example, *J.Med.Chem* paper "X" that contains assay data that defines compound "Y" as an inhibitor of protein "Z", proceeded to draw the chemical structure and link it with a public protein database identifier. In essence, both products use expert curators to extract such unstructured data and relationships from documents into databases on a large scale. The value, specificity challenges and quality control aspects associated with this process have been described in detail for WOMBAT which includes over 130,000 compounds with links to over 1,300 targets extracted from over 6,700 papers published in medicinal chemistry journals [16]. While there are differences in indexing, the largest difference to GVKBIO are scale and the inclusion of relationships extracted from patents as well as journals[17]. This has produced approximately 1.5 million compounds linked to over 2,000 sequences extracted from over 20,000 patents and journal articles.

These types of commercial databases fill a vital information gap for medicinal chemists because, while the provision of public connectivity between enzymes and cognate substrates has been addressed by ChEBI and KEGG (and include links into PubChem), the equivalent explicit links for bioactive compounds in general and drugs in particular, remains sparse. In some cases, indirect linkages can be made in PubChem via PubMed and MeSH but this is biased towards established bioactive compounds that are already indexed in MeSH. This problem has been recently and impressively tackled by DrugBank [18]. Although the compound coverage is modest compared to the commercial offerings described above, it is the first public initiative to explicitly link PubChem compound identifiers with the sequence identifiers of drug targets. A number of recent reviews have appeared that include merged compound counts and property analysis of chemical databases [15,19-21]. However, to our knowledge this is the first detailed comparison matrix between public and commercial compound databases focused on bioactive structures.

Box 1. Definitions of public, commercial, novel, proprietary and prior-art.

Public used to mean published compound structures from any source. However, because these are increasingly being captured in electronic repositories they are defined in practice by being either in public databases or commercial databases that extract their content from public primary sources. Consequently, we can define **novel** as a synthesised or proposed structure that has no identity match in those databases we use to define public.

Commercial refers to databases that we have purchased, imported the compound structures and where the identical data set is not freely accessible on the Web. There are hybrid access models, for example structures that are available in PubChem but a subscription is required to access the data links.

Proprietary implies private ownership that excludes other parties and may also be the subject of intellectual property law. It is often used as a synonym for **novel** but it is also transient because compounds designated as proprietary are merely presumed to be potentially patentable for therapeutic utility until the structure appears in, for example, one of the databases described in this review.

Prior-art, the opposite of proprietary, refers to the existence of not only the structure in public sources, but also inferences concerning bioactivity that could obviate or restrict a potential patent.

Box 2. Definitions of links and linking

The linking of information is a key concept in this review but is used in different contexts. The example used in the introduction is concept linking within the sentence “X (compound) inhibits Y (protein)”. This can be manually converted into a basic database of just four columns, the compound identifier, such as a IUPAC name or PubChem ID, the chemical structure, the protein name and a protein sequence identifier, such as a SwissProt or RefSeq ID. The utility for performing queries across these links can be increased by adding a document identifier for where the sentence was found, another column containing “inhibitor” to differentiate from “activators” and a protein function classifier, e.g. “protease”, thereby providing seven columns of links.

This covers two contexts of links. The third is the link, hyperlink, or Web link, used to connect the URL of a Web resource. Extending the example, each of the entries in the prototype database could be converted to a hyperlink to Web resources with a lookup facility. Thus, our prototype would then have outlinks. It could be converted into a Web resource in its own right by suggesting that relevant external resources modify their outlinks to point to it. These can be considered inlinks and thus become reciprocal. DrugBank is an example that includes outlinks to PubMed IDs. However, it currently lacks the reciprocal inlink i.e. DrugBank cannot yet be selected as a outlink source in PubChem (this situation is being addressed and reciprocal linking should soon become available [22]).

2.1 DATABASE DESCRIPTIONS

For an outline of most of the commercial databases used in this work we refer readers to a recent comprehensive review [23]. Included below are short descriptions of databases that have appeared since that review. The Web sites and, where available, publications that include some description of these resources, are listed in Table 2.

GVKBIO databases are populated with explicit relationships between compounds, assays and sequence identifiers that have been extracted from journals and patents on a large scale. We have included the Drug database of all FDA approved compounds, the MedChem database (700,000 entries) of compounds exclusively from medicinal chemistry journals and the target class databases divided into seven sections, GPCRs, proteases, kinases, ion-channels, NHRs, phosphatases and transporters. In total, these include 1.8 million entries merged between journals and patents. The GVKBIO databases are document-centric in that compound records have primary links to publications and patents [17].

PubChem is the NCBI public informatics backbone for the NIH Molecular Libraries Initiative focused on small molecules as systems biology probes and potential therapeutic agents [24]. It consists of PubChem Compound (unique structures) PubChem BioAssay (assay results) and PubChem Substance (structures supplied by depositors). The latest statistics, according to figures obtained from the website in November 2006, are 10.1 million compounds with 15.4 million links from 52 submitting sources. Of the compounds over 350,000 have been tested in 348 assays, over 10,000 are linked to protein 3D structures and over 70,000 to a publication via MeSH. This connectivity is integrated into the Entrez bioinformatic databases. The website includes comprehensive descriptions of data structures, content and mining tools.

DrugBank includes approximately 4300 small-molecule drug entries including FDA-approved drugs and experimental drugs. Over 6,000 protein target sequences are linked to these drug entries. Each DrugCard entry contains over 80 data fields, with half being devoted to drug and chemical data, and, the other half to target-centred information [18].

2.2 DATABASE SUBSETS

In addition to using complete compound collections we selected several subsets to provide insight into more specific areas of bioactive coverage. The PubChem subsets were extracted as PubChem identifiers by using the Entrez query and download facility. They were then converted to SDF Files by using the entire set downloaded from PubChem. The abbreviations used match the headings in Table 3.

- **PubChem Prous:** The *Drugs of the Future* Journal publishes monographs with information on new drug compounds in development. These were retrieved by searching “PubChem Compound” with the limit: “Prous Science Drugs of the Future” to give 3379 compound IDs.
- **PubChem PDB:** A complete set of small molecule ligands has been extracted from the Protein Data Bank, and deposited into PubChem [25]. These compounds were retrieved by using the limit of data source “SMID” to give 6128 compound IDs.
- **PubChem actives:** The PubChem BioAssay Database contains both purified enzyme assays and cell-based-bioactivity screens of chemical substances, although a number of the assays measure binding rather than activity modulation *per se*. The query limit “active in any bioassay” was used to retrieve 37349 compound IDs that should represent an *in vitro* bioactive subset.

- **PubChem pharmacol:** The MeSH index of PubMed includes the category “pharmacology” that is used for all drugs and exogenously administered chemical substances with effects on living tissues and organisms. It includes effects on physiological and biochemical processes as well as other pharmacological mechanisms of action. As most of the MeSH-defined compounds are linked into PubChem, the Entrez query limit “has pharmacological action” was used to retrieve 12038 compound IDs that should represent an *in vivo* bioactive subset.
- **GVKBIO journal:** Subset of GVKBIO compounds from the MedChem and Target databases selected by field code “journal” in the SD file.
- **GVKBIO patent:** Subset of GVKBIO compounds from the Target databases selected by field code “patent” in the SD file. These are selected predominantly as compounds from exemplified claims.
- **GVKBIO Drug Database:** FDA-approved and other drugs.
- **DrugBank_FDA:** FDA-approved small molecule drugs downloaded as set of SD files.
- **DrugBank_EXP:** Experimental drugs including unapproved, de-listed, illicit drugs, enzyme inhibitors and potential toxins. Downloaded as set of SDF files.
- **MDDR_launched:** Subset from MDDR selected by field code “launched” in the SD files.
- **ZINC_FDA:** Downloaded as 1217 SMILES files, date stamped as 1st of March 2005, as a subset provided by ZINC [26]. As ZINC compounds are included in PubChem it was not used as a separate set for the database comparisons.

2.3 DATABASE PROCESSING

All databases were downloaded as SD file format from their respective in-house, external registration system, or FTP site. The structures were converted into SMILES strings and all stereochemical information was removed in order to avoid differences between different conventions and allow a direct all-against-all database comparison [20]. We determined the number of unique structures in each database by the following four-step filtration procedure:

1. Normalisation of each molecule by removing small fragments, such as counterions, and neutralizing remaining charges.
2. Derivation of a canonical tautomer using LEATHERFACE, an in-house molecular editor based on SMARTS rules [27]
3. Generation of unique molecular hashcodes [28].
4. The retaining of unique structures by comparing the molecular hashcodes.

The same molecular hashcodes were subsequently used to identify the overlap between databases. All in-house scripts and codes used for this study are based on the OpenEye toolkit [29]. The effect of this processing on PubChem compound counts is presented in the Table 1 below

(Insert Table 1)

The download content and the results we obtained for all databases and subsets after this process are shown in Table 2. Analogous to the PubChem set shown in Table 1 the

numbers are consistently lower than those given for each of the sources, typically between 5% and 20%, although some sources do not explicitly specify compound numbers. Because structure matching is complex we do not claim our numbers as a “standard of truth” but this filtration process has proven robust and consistent for comparing in-house collections. The use of subsets provided internal controls and, in all cases, produced exact number matches i.e. the subsets added up the parent set. For very small percentage differences between independent data sources, it is not possible to discriminate between technical errors in small numbers of structure files and genuine unique content, unless a selection of these are manually inspected.

(Insert Table 2)

3.1 COMPARISON RESULTS

The comparison matrix is shown in Table 3.

(Insert Table 3)

We can review the matrix from left to right across the columns and down the rows in database order. The GVKBIO results include subsets split between journals and patents, with the GVKBIO drug database as a separate product (i.e. not a subset). At just under 1.5 million GVKBIO is divided between journals and patents at approximately 1:2 ratio, with an overlap of 89,000. There are also a small number of compounds in the Drug database that are not captured in the larger database. As expected from their conceptually similar strategies of capturing the explicit relationships between active compounds and their target sequence identifiers, the overlap between WOMBAT and GVKBIO is 93%. The PubChem overlap is proportionally lower, with 29% of GVKBIO represented in PubChem, but this is split evenly between journals and patents. As expected, PubChem includes 45% of compounds linked directly to journals in GVKBIO but only 22% of those that are linked to patents because PubChem has no content directly extracted from patents. However, the overlap shows that the number of PubChem compounds with potential patent claims is substantial at 238,000. The number of Prous *Drugs of the Future* journal-linked compounds in the PubChem subset is also substantial at 77% but not complete.

Moving across (row 1 in Table 3) we can see that GVKBIO includes just over half of the PDB ligands. In addition, it includes 25% of those compounds reported as being active in any of the screening data sets in PubChem and 70% of those with some kind of pharmacology link in PubChem via MeSH. The intersect between GVKBIO and PubChem pharmacology is explicable because many compounds whose activity against defined targets *in vitro* is published are also tested *in vivo* and therefore could be indexed in MeSH. The overlap between WOMBAT and GVKBIO shows that the former has captured over 7000 compounds not found in the latter, probably due to journal coverage differences. On the other hand only 29% of the Prous (*Drugs of the Future*) compounds are captured in WOMBAT. The indirect patent coverage shows the opposite trend to GVKBIO i.e. only 30,135 WOMBAT compounds match those extracted from patents by GVKBIO.

At just over 7 million compounds after filtration, PubChem (row 6 in Table 3) is by far the biggest set. In a comparison not included in the table we established that 48% of its content overlapped with the chemical suppliers included in the ChemNavigator iResearch

Library (July 2005 release) that includes over 13 million available screening compounds and reagents [30]. The inference, supported by the primary and secondary (e.g. via ZINC) submitting sources, is a “vendor push” into PubChem. While this could dilute out the bioactive component, today’s vendor compound could be tomorrow’s bioactive. The Entrez query “Tested in any BioAssay” shows that only just over 3% of PubChem has *in vitro* assay results in the system. Notwithstanding, PubChem shows the largest coverage of every database in Fig.3, with the exception of WOMBAT, as already described, of which PubChem covers some 3,000 less compounds than GVKBIO. Moving across row 6 it covers, 92% of GVKBIO Drugs (even just exceeding those covered by GVKBIO), 95% of MDDR launched, 46% of DNP, 42% of MDDR, 93% of CMC and 95% of BioPrint. While it also contains 98% of ZINC FDA and 92% of DrugBank both these databases specifically link to PubChem.

Moving down (column 7 in Table 3) shows the overlaps between WOMBAT, GVKBIO and the Prous (*Drugs of the Future*) compounds in PubChem. This highlights key differences in the databases because, while the linkage between the compounds and the documents are explicit in PubChem, this is not reciprocal. This means there are neither links from articles out to PubChem IDs nor between compounds and sequence identifiers of the targets that are, in most cases, identifiable in the context of the document. For 77% of the compounds, GVKBIO has filled this gap, even though *Drugs of the Future* is not currently a source journal, by manually curating the primary literature for extracting the compound-sequence links rather than these particular review articles. The PDB ligand has low coverage in all other databases, with the exception of DrugBank experimental drugs, 41% of which are in PDB. The PubChem actives show even lower overlap with the other sources. This uniqueness could be because this raw data is not captured by commercial sources. It is likely to increase considerably as the public HTS screening centers deposit more data into PubChem. As noted above, only GVKBIO overlaps with a significant proportion of the PubChem pharmacology subset. After PubChem, the DNP and MDDR databases show the largest proportion of unique content, although without further analysis it is not possible to discern if the overlap between MDDR with GVKBIO and PubChem covers the same or a different set of compounds.

The collection of approved drugs is a key compound set from many perspectives, not least of which is defining relationships between those compounds and sequence identifiers for their targets. A number of databases are either based on, or claim to include, this subset of bioactives. In fact Table 3 shows that only CMC, PubChem, GVKBIO journals and BIOPRINT have $\geq 90\%$ of the 1200 compounds, even though those such as MDDR launched, WOMBAT, GVKBIO Drugs and DrugBank might be expected to give a more extensive overlap. Update frequency will affect relative coverage and this may be the case with DrugBank where the 102 experimental drugs that overlap with the FDA set could now be re-classified as approved.

3.2 MERGED AND UNIQUE CONTENT

A limitation of the data in Table 3, already alluded to in the MDDR analysis, is that unique content can only be inferred where the sum of overlaps is below the total. True unique content therefore needs to be defined in terms of a Venn-type series of $B+C+D-A$, $A+C+D-B$ and so on. While this would be an extensive undertaking in its entirety, we have merged all the databases in Table 3. As implied from the discussion above, only the PubChem subsets are included rather than the entire database, thereby merging all these sources of bioactive compounds to give 1,976,273. When this is filtered, the unique content reduces to 1,741,392. This relatively small redundancy collapse of 234,881 (11%) indicates substantial unique

content overall, once again emphasising the complementarity. We investigated this further by producing full Venn-type overlaps for PubChem, GVKBIO and WOMBAT, shown in Fig. 1

(insert Figure 1)

The Venn-type display shows a number of details that provide additional information. The three-way overlap is 86,143. This shows the positive aspect of redundancy where three-fold independent confirmation defines a high-value consensus bioactive subset. Thus, these compounds not only have duplicate, manually confirmed, explicit publication-compound-sequence identifier links, but, also outlink to many additional information sources via the PubChem system. As discussed above, the vendor push would contribute to the 93% of PubChem that lies outside the other two databases but this would also include some bioactive content for which the connection between biological effect, *in vitro* activity and target sequence identifier is not explicit in the documents covered so far by the other two databases. Another conspicuous feature of Fig. 1 is that for over a million compounds (68%) GVKBIO is unique. This has utility that can be seen differently depending on the context. For a pharmaceutical company, this set is valuable for determining prior-art, especially as part of the 1.7 million merged set described above (together with the obvious addition of PubChem). From a scientific point of view, this set of compound-sequence identifier links fills a big information gap that text-mining does not yet have the precision to fill [31,32]. However, in our experience it is necessary to be circumspect about bioactivity claims from patent-only specified relationships, even if the GVKBIO curators confine these to exemplified compounds. Thus, the overlap set between PubChem and GVKBIO of nearly 354,000 may have a higher authenticity and verification rate for bioactivity than those outside it but not as high as we might expect from the three-way set. The figure also shows that over 3% of WOMBAT is also unique in this comparison.

4. CONCLUSIONS

The results we have reviewed from pairwise and Venn-type comparisons demonstrate the powerful complementarity of commercial and public sources for coverage of bioactive chemical space. The degree of non-overlap i.e. few sets are entirely nested within others, was unexpected and indicates substantial unique content. We would like to emphasise that our descriptions of relative coverage should neither be interpreted as a criticism, nor as an endorsement, for any particular database. Notwithstanding, some differences we have reported, for example in the known drugs sets, could be easier to interpret if more clarity was provided on the technicalities of data selection and extraction from what are, ostensibly, the same primary public sources. Database providers could also consider performing the kind of analysis we have shown here as part of their efforts to demonstrate utility.

The commercial products included in this work are evaluated in-house not only by content coverage *per se* but also by cost, internal complementarity, link richness, technical compatibilities, user access models, learning curves, update frequencies and data mining options they provide. Public databases are assessed in a different context because funding bodies support them as an invaluable service to the scientific community, but their utility is examined and compared by the same standards. Consequently, commercial databases now have to compete with the information-rich links, advanced data structures and sophisticated query options of public databases, of which Entrez and the DrugBank Data Extractor are good

examples. These also have the increasingly important advantage of reciprocal connectivity with other Web databases.

Within AstraZeneca, this approach of selecting bioactive chemical sources based on complementarity has proved valuable. Other companies are also exploiting this type of information because it facilitates choices for chemotype starting points from which to push towards novel drug-like compounds unencumbered by prior-art [33]. Success in this chemical “space race” therefore has to include comparison of proprietary in-house screening compounds and activity results obtained from them against the available bioactive collections. Clearly this has pragmatic limitations, not the least of which would be the expense of subscribing to all the commercial offerings in this area. In addition, there are a number of on-line-only resources for which their usual access and licensing models preclude inclusion of their content in the type of comparative analysis we have described.

Our analysis makes clear that the comparison of public sources with commercial collections is increasingly important. The wider implications of the proliferation of public sources are outside the scope of this review, but a key scientific aspect is that they are extending beyond the boundaries of the traditional medicinal chemistry focus on the development of drug candidates. Public collections now have much broader utilities and scope. These include the provision of tool compounds for systems biology, imaging reagents, and chemogenomic probes as well as cognate enzyme substrates and receptor ligands. They are also pushing to define chemical space for the lipidome, glycome and metabolomes of model organisms [34-36]. These wider areas of chemical biology and biochemistry are largely devoid of commercial equivalents, with the possible exception of natural product databases that can be considered as subsets of exotic metabolomes.

Table 1. Effect of processing on PubChem compound counts

PubChem	Count	Comments
Download	10,062,600	After SDF to SMILES conversion
Unique structures with stereochemistry	8,846,124	canonical tautomers
Unique structures without stereochemistry	7,268,193	canonical tautomers

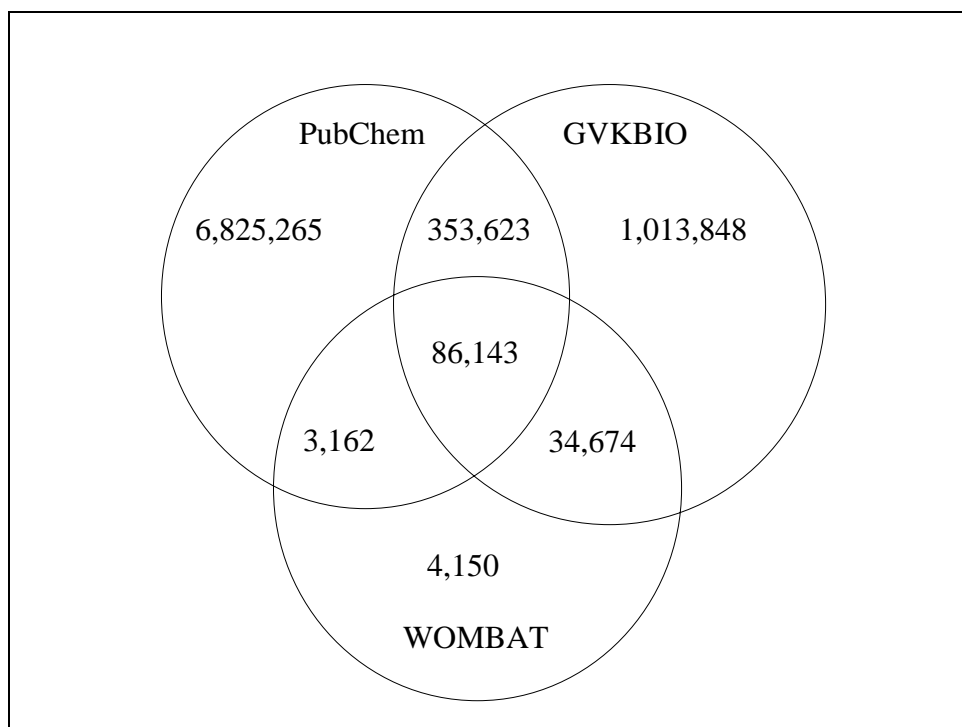
Table 2. Database Information and Unique Compound Content

Database	Short Name	Ref.	Given compounds or records	Unique compounds	Version / Download time stamp	URL
GVKBIO			2,742,213	1,439,678	Jul 2006	www.gvkbio.com/informatics/dbprod.htm
World of Molecular Bioativity	WOMBAT	[16]	154,234	128,120	2006.1	sunsetmolecular.com/products/?id=4
DrugBank		[18]	4,261	3,755	Jul 2006	redpoll.pharmacy.ualberta.ca/drugbank/index.html
PubChem			10,062,600	7,268,193	Nov 2006	pubchem.ncbi.nlm.nih.gov/
Dictionary of Natural Products	DNP		178,347	131,831	Apr 2006	www.ramex.com/cr/cr-dict0.html
MDL Drug Data Report	MDDR	[37]	169,242	159,867	2006.1	www.mdl.com/products/knowledge/index.jsp
MDL Comprehensive Medicinal Chemistry	CMC		8,757	8,189	2005.1	www.mdl.com/products/knowledge/index.jsp
BioPrint	BIOPRINT	[38]	2,490	2,437	Oct 2006	www.cerep.fr/cerep/users/pages/Collaborations/BioPrint.asp

Table 3. Compound Overlap Matches Between Databases and their Subsets

	GVKBio	GVKBio Journals	GVKBio Patents	GVKBio Drug db	WOMBAT	PubChem	PubChem Prous	PubChem PDB	PubChem actives	PubChem pharmacol	DrugBank	DrugBank small mol	DrugBank exp drug	DNP	MDDR	MDDR launch	CMC	BioPrint	ZINC FDA
GVKBio	1 488 288 542 858 1 034 548 1 663	120 817	439 766	2 563		2 628	9 347	4 221	2 415	958	1 491	11 662	64 733	971	5 205	2 024	1 089		
GVKBio Journals	542 858 89 118	1 637	119 875	245 735	2 424	2 522	9 102	4 090	2 353	945	1 442	10 990	28 887	949	5 023	1 990	1 078		
GVKBio Patents		1 034 548 977	30 135	237 643	1 215	1 052	2 277	2 105	1 234	680	581	1 968	46 712	646	2 086	1 192	633		
GVKBio DrugDatabase			1 933	683	1 795	733	260	647	1 402	966	891	127	288	940	870	1 723	1 191	840	
WOMBAT				128 120	89 305	995	1 114	3 545	1 568	1 107	493	643	1 719	11 601	446	1 471	913	437	
PubChem					7 268	1933 318	5 626	35 671	6 070	3 445	1 001	2 477	61 266	68 178	1 063	7 678	2 328	1 180	
PubChem Prous							3 318	236	541	1 374	543	448	118	425	1 875	734	1 792	637	315
PubChem PDB								5 626	538	767	2 383	186	2 237	1 032	661	137	516	287	185
PubChem actives									35 671	1 690	721	459	290	1 999	1 323	318	1 417	951	464
PubChem pharmacol										6 070	1 209	881	374	1 182	1 507	862	2722	1 555	932
DrugBank											3 723	1 018	2 737	819	869	571	1 172	907	639
DrugBank small mol												1 018	65	134	587	547	930	814	573
DrugBank exp drug													2 737	691	314	57	283	143	102
DNP														131 831	2 602	105	1129	392	232
MDDR															159 867	1 118	2185	867	468
MDDR launched																1 118	1 049	655	423
CMC																	8 189	1 669	1 125
BioPrint																		2 437	819
ZINC FDA																			1 200

Figure 1. Venn Diagram of the Complete Overlaps Between PubChem, GVKBIO and WOMBAT. Compound totals, after filtration as described, are 7,268,193, 1,488,288, and 128,129, respectively.



REFERENCES

- [1] Wishart, D. S. Bioinformatics in drug development and assessment. *Drug Metabolism Reviews* **2005**, *37*, 279-310.
- [2] Murray-Rust, P.; Mitchell, J. B.; Rzepa, H. S. Chemistry in bioinformatics. *BMC Bioinformatics* **2005**, *6*, 141.
- [3] Bradley, D. Public molecules: small, but perfectly formed. *Nature Reviews Drug Discovery* **2004**, *3*, 988-989.
- [4] Brooksbank, C.; Cameron, G.; Thornton, J. The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Research* **2005**, *33*, D46-53.
- [5] Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2006**, *34*, D173-180.
- [6] Wilbur, W. J.; Hazard, G. F., Jr.; Divita, G.; Mork, J. G.; Aronson, A. R. et al. Analysis of biomedical text for chemical names: a comparison of three methods. *Proceedings / AMIA 1999, Annual Symposium.*, 176-180.
- [7] Westbrook, J.; Feng, Z.; Jain, S.; Bhat, T. N.; Thanki, N. et al. The Protein Data Bank: unifying the archive. *Nucleic Acids Research* **2002**, *30*, 245-248.
- [8] Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. *Journal of Medicinal Chemistry* **2002**, *45*, 818-840.
- [9] Strausberg, R. L.; Schreiber, S. L. From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science* **2003**, *300*, 294-295.
- [10] Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: data management and interface design. *Bioinformatics* **2002**, *18*, 130-139.
- [11] Baker, M. Open-access chemistry databases evolving slowly but not surely. *Nature Reviews. Drug Discovery* **2006**, *5*, 707-708.
- [12] Parker, C. N.; Shamu, C. E.; Kraybill, B.; Austin, C. P.; Bajorath, J. Measure, mine, model, and manipulate: the future for HTS and cheminformatics? *Drug Discov Today*. **2006**, *11*, 863-865.
- [13] Geldenhuys, W. J.; Gaasch, K. E.; Watson, M.; Allen, D. D.; Van der Schyf, C. J. Optimizing the use of open-source software applications in drug discovery. *Drug Discovery Today* **2006**, *11*, 127-132.
- [14] Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H. et al. The Blue Obelisk-interopability in chemical informatics. *Journal of Chemical Information & Modeling* **2006**, *46*, 991-998.
- [15] Richard, A. M.; Gold, L. S.; Nicklaus, M. C. Chemical structure indexing of toxicity data on the internet: moving toward a flat world. *Current Opinion in Drug Discovery & Development* **2006**, *9*, 314-325.
- [16] Olah, M.; Rad, R.; Mracec, M.; Ostopovici, L.; Bora, A. et al. WOMBAT and WOMBAT-PK: bioactivity databases for lead and lead discovery. *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*; Wiley-VCH Verlag GmbH: Weinheim, 2007; pp in press.
- [17] GVKBIO <http://www.gvkbio.com>.
- [18] Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* **2006**, *34*, D668-672.

- [19] Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol Divers.* **2006**, *10*, 389-403.
- [20] Engels, M. F. M.; Gibbs, A. C.; Jaeger, E. P.; Verbinnen, D.; Lobanov, V. S. et al. A Cluster-Based Strategy for Assessing the Overlap between Large Chemical Libraries and Its Application to a Recent Acquisition. *J. Chem. Inf. Model.* **2006**.
- [21] Muresan, S.; Sadowski, J. "In-house likeness": comparison of large compound collections using artificial neural networks. *Journal of Chemical Information & Modeling* **2005**, *45*, 888-893.
- [22] Wishart, D. S. personal communication.
- [23] Jonsdottir, S. O.; Jorgensen, F. S.; Brunak, S. Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics* **2005**, *21*, 2145-2160.
- [24] Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138-1139.
- [25] Snyder, K. A.; Feldman, H. J.; Dumontier, M.; Salama, J. J.; Hogue, C. W. Domain-based small molecule binding site annotation. *BMC Bioinformatics* **2006**, *7*, 152.
- [26] ZINC <http://blaster.docking.org/zinc/>.
- [27] Oprea, T. I. Chemoinformatics in lead discovery. *Chemoinformatics in Drug Discovery*; Wiley-VCH Verlag GmbH: Weinheim, 2005; pp 25-38.
- [28] Nilakantan, R.; Bauman, N.; Haraki, K. S. Database diversity assessment: new ideas, concepts, and tools. *Journal of Computer-Aided Molecular Design* **1997**, *11*, 447-452.
- [29] OpenEye Scientific Software www.eyesopen.com.
- [30] ChemNavigator <http://www.chemnavigator.com/>.
- [31] Banville, D. L. Mining chemical structural information from the drug literature. *Drug Discovery Today* **2006**, *11*, 35-42.
- [32] Rodriguez-Esteban, R.; Iossifov, I.; Rzhetsky, A. Imitating Manual Curation of Text-Mined Facts in Biomedicine. *PLoS Comput Biol.* **2006**, *2*(9), 24-20.
- [33] Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nature Biotechnology* **2006**, *24*, 805-815.
- [34] Cotter, D.; Maer, A.; Guda, C.; Saunders, B.; Subramaniam, S. LMPD: LIPID MAPS proteome database. *Nucleic Acids Research* **2006**, *34*, 507-510.
- [35] Hashimoto, K.; Goto, S.; Kawano, S.; Aoki-Kinoshita, K. F.; Ueda, N. et al. KEGG as a glycome informatics resource. *Glycobiology* **2006**, *16*, 63-70.
- [36] Human Metabolite Database (HMD) <http://www.hmdb.ca/>.
- [37] Sheridan, R. P.; Shpungin, J. Calculating similarities between biological activities in the MDL Drug Data Report database. *Journal of Chemical Information & Computer Sciences* **2004**, *44*, 727-740.
- [38] Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B. et al. Predicting ADME properties and side effects: the BioPrint approach. *Curr Opin Drug Discov Devel.* **2003**, *6*, 470-480.