Journal of
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

Analysis of Commercial and Public Bioactivity Databases

Pekka Tiikkainen*,[†] and Lutz Franke[†]

[†]Merz Pharmaceuticals GmbH, Eckenheimer Landstrasse 100, 60318 Frankfurt am Main, Germany

S *Supporting Information*

ABSTRACT: Activity data for small molecules are invaluable in chemoinformatics. Various bioactivity databases exist containing detailed information of target proteins and quantitative binding data for small molecules extracted from journals and patents. In the current work, we have merged several public and commercial bioactivity databases into one bioactivity metabase. The molecular presentation, target information, and activity data of the vendor databases were standardized. The main motivation of the work was to create a single relational database which allows fast and simple data retrieval by in-house scientists. Second, we wanted to know the amount of overlap between databases by commercial and public vendors to see whether the former contain data complementing the latter. Third, we quantified the degree of inconsistency between data sources by comparing data points derived from the same scientific article cited by more than one vendor. We found that each data source contains unique data which is due to different scientific articles cited by the vendors. When comparing data derived from the same article we found that inconsistencies between the vendors are common. In conclusion, using databases of different vendors is still useful since the data overlap is not complete. It should be noted that this can be partially explained by the inconsistencies and errors in the source data.

INTRODUCTION

Databases containing associations of small molecules with protein targets are vital in drug discovery. Virtual screening campaigns depend on these databases for active molecules used both as starting points in drug discovery programs and in evaluation of virtual screening performance. Researches interested in polypharmacology use the databases to find targets for a compound of interest.[1] The database content has been usually collected manually from scientific articles and patents. This is a costly strategy meaning that most large databases have been fee-based to cover the costs. Public databases have been relatively small concentrating on specific subsets of small molecules and targets (e.g., Drugbank[2]). One of the greatest changes in the situation was the release of the ChEMBL[3,4] database with its close to three million data points in the public domain in early 2010.[5]

With multiple databases available, combining them into a single data resource makes sense to make data retrieval and upkeep faster and simpler. Paolini et al.[6] merged bioactivity data from commercial and in-house databases and used the resource to enumerate protein targets and the degree of their promiscuity, i.e. the number of small molecules they share. In a similar effort Southan and co-workers[7] concentrated in the overlap of small molecules of six distinct databases and found that while the databases often overlap substantially each database also contains unique data. They suggest that commercial and public databases are complementary and should be used alongside each other. Taboureau and Nielsen et al.[8] have launched a database titled "ChemProt" which is publicly available through a Web-based user interface.[9] The resource combines data from various databases. One of these is the commercial WOMBAT database,[10] but here only three WOMBAT activity data points are shown per query (Tudor Oprea, personal communication).

In the current work, we have built an integrated data source of bioactivity data (for clarity, referred to as "Metabase" hereon) collected from both commercial and public sources. Our primary motivation was to collect data available to us into a standardized source that is readily available to in-house scientists. As in the work mentioned above, one of our study objectives was to quantify the overlap and complementarity of the underlying data sources (termed "vendor databases" hereon). The overlap of commercial databases with public ones was especially interesting. The second objective was to use standardized citations to investigate the consistency between vendor databases, i.e. whether all databases citing a particular journal article had been able to extract exactly the same information. Additionally, we have quantified the bias the vendor databases have on different target classes and the relative popularity of the target classes over the years.

MATERIALS AND METHODS

Metabase Structure. The Metabase has been implemented as a MySQL relational database with different entities such as molecules, targets, and activities separated into their respective tables. The tables are linked with foreign key constraints to ensure data integrity.

Vendor Databases. *PubChem:*[11] Actives from 561 confirmatory PubChem bioassays (see Table S1 in the Supporting Information) with a specified protein target and quantitative dose−response assay data were included. Assay information and small molecule structures were downloaded from the PubChem FTP server.[12] *WOMBAT:* Version 2011.01 of this commercial database available from Subset Molecular Inc. was used in the analysis.[10,13] The database is distributed as MDL/ISIS files which were exported as

Special Issue: 2011 Noordwijkerhout Cheminformatics

Received: July 8, 2011
Published: December 06, 2011

RD files readable by Pipeline Pilot.[14] *ChEMBL*:[3] The MySQL distribution of the database (Build 09) was downloaded from the ChEMBL FTP server[4] and installed locally. Fields required in the Metabase were extracted with a single SQL query. *Evolvus*: this commercial database by Evolvus[15] contains bioactivity data manually collected from both patents and journal articles. The curation effort has centered on targets important in drug discovery: GPCRs, kinases, ion channels, nuclear receptors, proteases, and phosphatases. These data were provided to us as a collection of RD files. *Ki Database*:[16] the text file containing the database (version 21_01_11) was downloaded from the vendor home page.[17]

We would like to point out that we have excluded BindingDB,[19] a public bioactivity database, from this analysis. Much of its content has been extracted directly from ChEMBL and PubChem, and therefore these data are not a result of an independent curation effort.[20] Including the database would have confounded figures for data overlap and uniqueness reported below.

Preparation of Vendor Databases. Both commercial (WOMBAT and Evolvus) and public (ChEMBL, PubChem and Ki Database) bioactivity databases were integrated into the Metabase. Across the vendors, the data format, variable names, and types of data delivered are different. For each vendor, a custom Pipeline Pilot protocol was designed that picks and renames important fields and writes the data into an SD file with one record for each activity. This file is compatible with a downstream Pipeline Pilot protocol performing the actual data incorporation into the Metabase.

As different tautomeric and charge isoforms of the same small molecule are provided by different vendors, the molecular structures are standardized in the incorporation step. Standard components in Pipeline Pilot 8.0[14] were used in the following standardization steps: only the largest fragment in the molecule record was kept, stereochemistry and charges were standardized using the "Standardize Molecule" component. Furthermore bases were deprotonated with the "Deprotonate Bases" component and acids protonated with the "Protonate Acids" component. Next the canonical tautomer was determined by enumerating all possible tautomers ("Enumerate Tautomers" component) and picking the one with the largest tautomer score given by the component. If more than one tautomer had the same score, the canonical tautomer with the alphabetically greatest canonical smiles string was picked. Molecules were kept as the stereoisomers they were provided by the vendor. Finally an InChI key[21] (version 1.03) was calculated for each molecule with Pipeline Pilot's "Molecule to InChI" component to avoid redundancy. All these steps led to a small drop in the number of distinct molecular entities (Table S2 in the Supporting Information). Each InCHI key is associated with an internal Molecule ID used to refer to the structure across the Metabase. Whenever Uniprot IDs annotating molecular targets were given in the activity data, they were used to link the given activity with a list of all Uniprot IDs contained in Metabase. Finally, the protocol checks whether an activity from the same vendor with the same combination of small molecule, target, citation, and other fields already exists in the Metabase to avoid data duplication.

Unification of Activity Data Types and Units. Bioactivity data are provided in a plethora of activity types and concentration units. For example, IC$_{50}$ values are given either as linear values or as negative logarithms of the concentration (pIC50). To allow for comparison of data, bioactivity value representation was unified to linear values with the concentration expressed as micromoles per liter. Unification was done only for activity values derived from dose−response measurements, while single point inhibition or activation data were ignored. Activity values less than 1 fM and greater than 1 M were further excluded as such extreme values are likely to result from errors in the source document itself or the data extraction process.

Table 1. General Statistics of the Metabase

statistic	count
number of unique molecules	2,477,290
...of which associated with at least one activity	844,553
distinct activities	2,041,996
number of distinct targets	
...Uniprot IDs	5,857
...recommended protein names	3,446
articles cited	43,174
patents cited	27,018

In the following an activity is defined as a unique combination of the small molecule structure, Uniprot ID and unified activity value, unit, type, and relation. All variables had to be defined for an activity to be taken into analysis. When comparing activities, the activity values were rounded to the two most significant digits to negate the rounding errors that arose from conversion of logarithmic values into linear ones.

Protein Targets. Ligand protein targets are identified by their respective Uniprot IDs.[22] Uniprot accessions are species-specific, e.g. P28222 is the human 5-hydroxytryptamine receptor 1B. Ki Database and PubChem do not contain Uniprot accessions, and these had to be looked-up for the two databases. The former provides a Unigene code[23] and a species name for each protein target, and these were converted into a Uniprot accession with a look-up table. On the other hand, PubChem contains GenInfo Identifiers (GI numbers)[24] which were converted into Uniprot accessions with a look-up table. Both look-up tables were obtained from the Uniprot FTP server.[25]

To allow the retrieval of ligands tested against any ortholog of a protein, Uniprot accessions were further linked either to a Uniprot recommended protein name (manually curated data) or to a submitted protein name (sequences still waiting for manual curation).[26] For example, the recommended protein name "5-hydroxytryptamine receptor 1B" is a common term for 17 orthologs of the protein in different species. In total the metabase contains activities for 5,857 unique Uniprot IDs (2,423 of these human) corresponding to 3,446 distinct recommended or submitted protein names (Table 1).

Mapping the protein targets to protein families allows a more general view of activity data distribution. We concentrated on five major therapeutically important protein classes: G-protein coupled receptors (GPCRs), ion channels, nuclear receptors, enzymes, and transporters. The hierarchies for the first three classes were taken from the IUPHAR database,[27,28] while the enzyme hierarchy (EC codes) was downloaded

Table 2. Unique Small Molecule Structures and Activities by Vendor

vendor	total molecules	molecules unique to the vendor	% unique molecules	total activities	activities unique to the vendor	% unique activities
ChEMBL	585,225	327,651	56.0	904,841	660,675	73.0
Evolvus	1,899,413	1,715,497	90.3	837,628	701,355	83.7
Ki Database	3,887	959	24.7	30,739	26,268	85.5
PubChem	89,993	85,240	94.7	200,889	200,839	99.9
WOMBAT	251,240	71,888	28.6	378,743	193,301	51.0

Table 3. Distribution of Activities Across Five Major Protein Classes[a]

protein class	ChEMBL	WOMBAT	PubChem	Evolvus	Ki Database	all vendors
enzymes	349,821 (38.7%)	165,291 (43.6%)	59,700 (29.7%)	267,565 (31.9%)	2,562 (8.3%)	740,635 (36.3%)
GPCR	275,928 (30.5%)	128,122 (33.8%)	8,653 (4.3%)	375,571 (44.8%)	24,295 (79.0%)	667,692 (32.7%)
ion channel	138,925 (15.4%)	25,139 (6.6%)	1,873 (0.9%)	69,835 (8.3%)	1,256 (4.1%)	215,111 (10.5%)
nuclear receptor	31,107 (3.4%)	16,460 (4.4%)	29,073 (14.5%)	36,918 (4.4%)	151 (0.5%)	95,972 (4.7%)
transporters	43,543 (4.8%)	24,845 (6.6%)	8,160 (4.1%)	39,978 (4.8%)	2,579 (8.4%)	103,161 (5.1%)
others	78,399 (8.7%)	27,260 (7.2%)	93,044 (46.3%)	54,613 (6.5%)	263 (0.9%)	219,425 (10.7%)

[a] Each cell contains the absolute number of activities. Percentage points in parentheses give the share of the vendor's activities in the given class. The value in the 'all vendors' column is less than the sum of values for individual vendors of the corresponding row since data overlap is taken in account. Accordingly, the sum of activities for each column can be higher than the activity counts given in Table 2 since some targets are assigned to more than one protein class. If a target appears more than once within a protein class or a subclass, its activities are counted only once. A more detailed table is available in Table S4 in the Supporting Information.

from the Enzyme nomenclature database.[29] The transporter classification was taken from the Transporter Classification Database (TCDB).[30,31] Ion channel families in the TCDB (family identifiers 1.A.1, 1.A.10, 1.A.2, 1.A.4, 1.A.7, and 1.A.9) were excluded to avoid duplication with IUPHAR's ion channels. It should be noted that the EC codes group enzymes by the type of a chemical reaction they catalyze and not by their sequence similarity and therefore do not strictly speaking correspond to protein families. These five classes cover 76.4% (2,632 out of 3,446) of the targets in the Metabase and 89.3% (1,822,571 out of 2,041,996) of all activities.

Unification of Journal Article Citations. Four of the vendor databases (ChEMBL, Evolvus, Ki Database, and WOMBAT) contain citations to journal articles. In Metabase, each article is identified as a unique triplet of an internal journal identifier, volume, and the first page of the article. Ki Database contains Pubmed identifiers[32] for its citations. These were converted into the triplet with a look-up table. For the remaining three vendor databases, the citation was parsed from a text field. Different abbreviations were given for the journals which were converted into the internal journal identifier with a manually built look-up table. In total, Metabase contains 43,174 journal article citations. Vendor database specific article counts are available in Table S3 in the Supporting Information.

RESULTS AND DISCUSSION

Metabase Statistics and Vendor Uniqueness. General statistics of the Metabase are given in Table 1. It is noteworthy that only 34.1% of the molecules (844,553 out of 2,477,290) are associated with a unified activity. The remaining molecules have nonstandardized target information (missing Uniprot identifiers or the target is a cell line, etc.), have only nondose response data activity associated with them, or the associated activity value is outside the range of realistic values (1 fM to 1 M).

The vendor databases vary widely in their data uniqueness. Table 2 shows the total count of small molecules and activities for each vendor plus the ratio of molecules and activities mentioned solely by this specific vendor. The vendor database with the most unique content is PubChem providing data of screening campaigns, while other vendors curate their data from scientific journals and patents. Another distinctively unique data source is Evolvus which is the only one of the considered vendors using patents as a data source. The patents are a complementary data source, but only 9,729 out of the 27,018 patents (36.0%) contain any activities with Uniprot IDs and unified activity values. The corresponding ratio is 22,113 out of 43,174 (51.2%) for journal articles. The average patent has 22.3 activities, and the average article has 20.0 activities.

The commercial vendor databases (Evolvus and WOMBAT) are the only source for 1,804,485 out of 2,477,290 molecules (72.8%) which drops to 53.1% (448,799 out of 844,553) if only molecules associated with at least one activity are considered. Out of the 2,041,996 activities, 909,841 (44.6%) are exclusively from the two commercial vendors. Clearly the commercial databases are still a valuable addition to publicly available data despite the rapid growth of the latter in the recent years. This was also one of the conclusions drawn by Southan and co-workers[7] who also noticed that the complementarity between the commercial and public databases had been growing over the three-year time span analyzed in their work.

A detailed overlap of molecules and activities across all combinations of vendors is given in Table S3 in the Supporting Information.

Target Family Analysis. Out of the five protein groups, enzymes have the most activities (740,635) followed by GPCRs (667,692), ion channels (215,111), transporters (without ion channels) (103,161), and nuclear receptors (95,972) (Table 3).

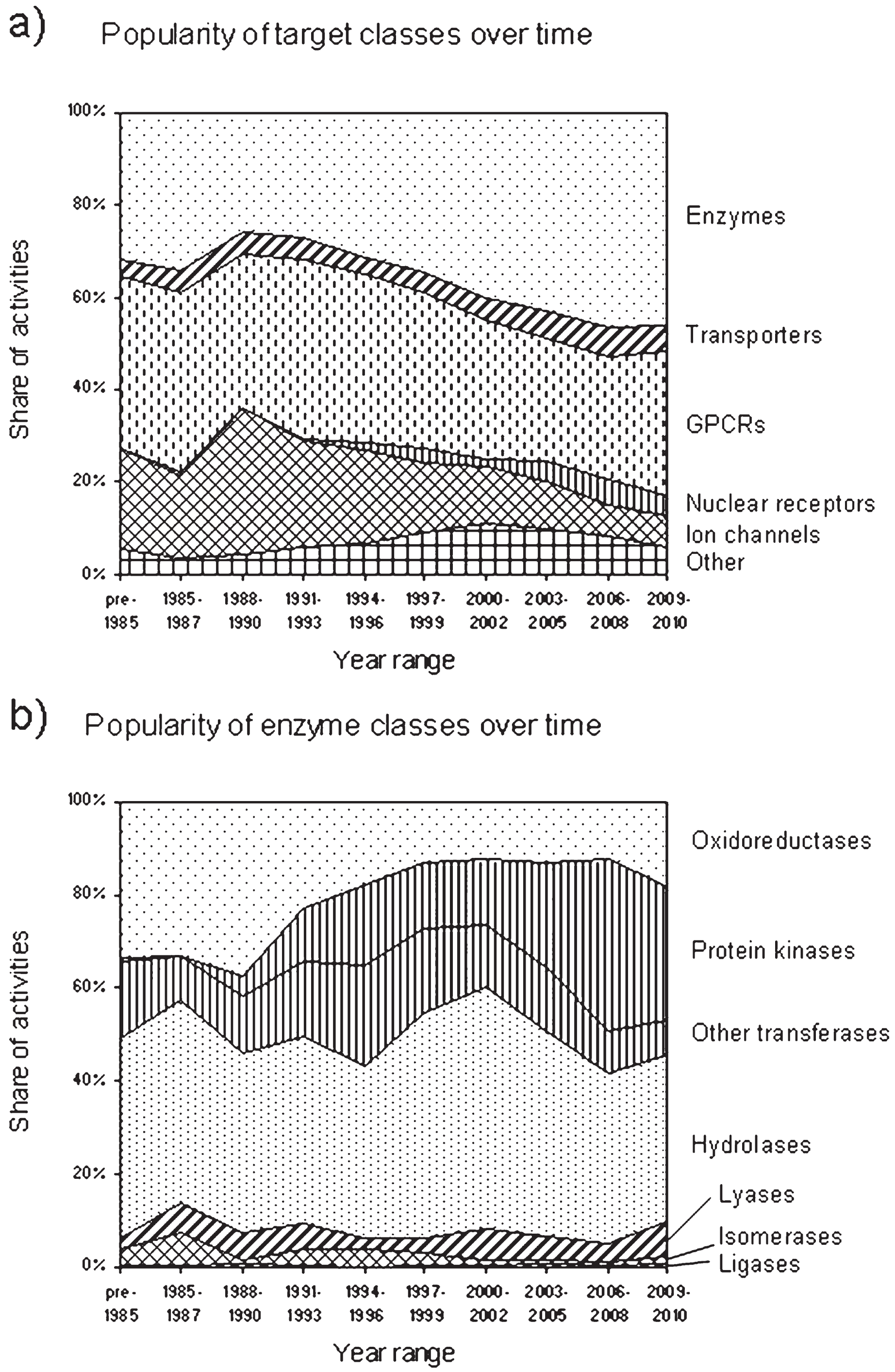


Figure 1. Relative shares of all activities assigned to (a) different protein classes and (b) enzyme subclasses in three-year ranges. Only activities extracted from journal articles are considered. Noteworthy points are the relative increase in activity data published for enzymes and decrease in the number of publications for ion channels. Within enzymes, the dramatic increase in activities for protein kinases (EC codes 2.7.10.- and 2.7.11.-) is especially noteworthy.

The numbers mirror the frequency of drug targets[33] only partially. For example, GPCRs are the second largest class in Table 3, but the most frequent target of launched drugs. Nuclear receptors are seriously under represented in the Metabase with 4.6% of all activities although 13% of launched drugs target one.[33]

Distribution of activities across protein classes can vary greatly between vendors. For instance, the Ki Database contains relatively more activities for GPCRs (79.0%) than the Metabase in total (32.7%). Another example is PubChem's preference for nuclear receptors and miscellaneous targets not included in the five major target classes (row "Others" in Table 3) with 14.5% and 46.3% of activities, respectively. On the other hand PubChem contains little data for ion channels and GPCRs − both classical drug target groups. Select PubChem bioassays have been incorporated in ChEMBL

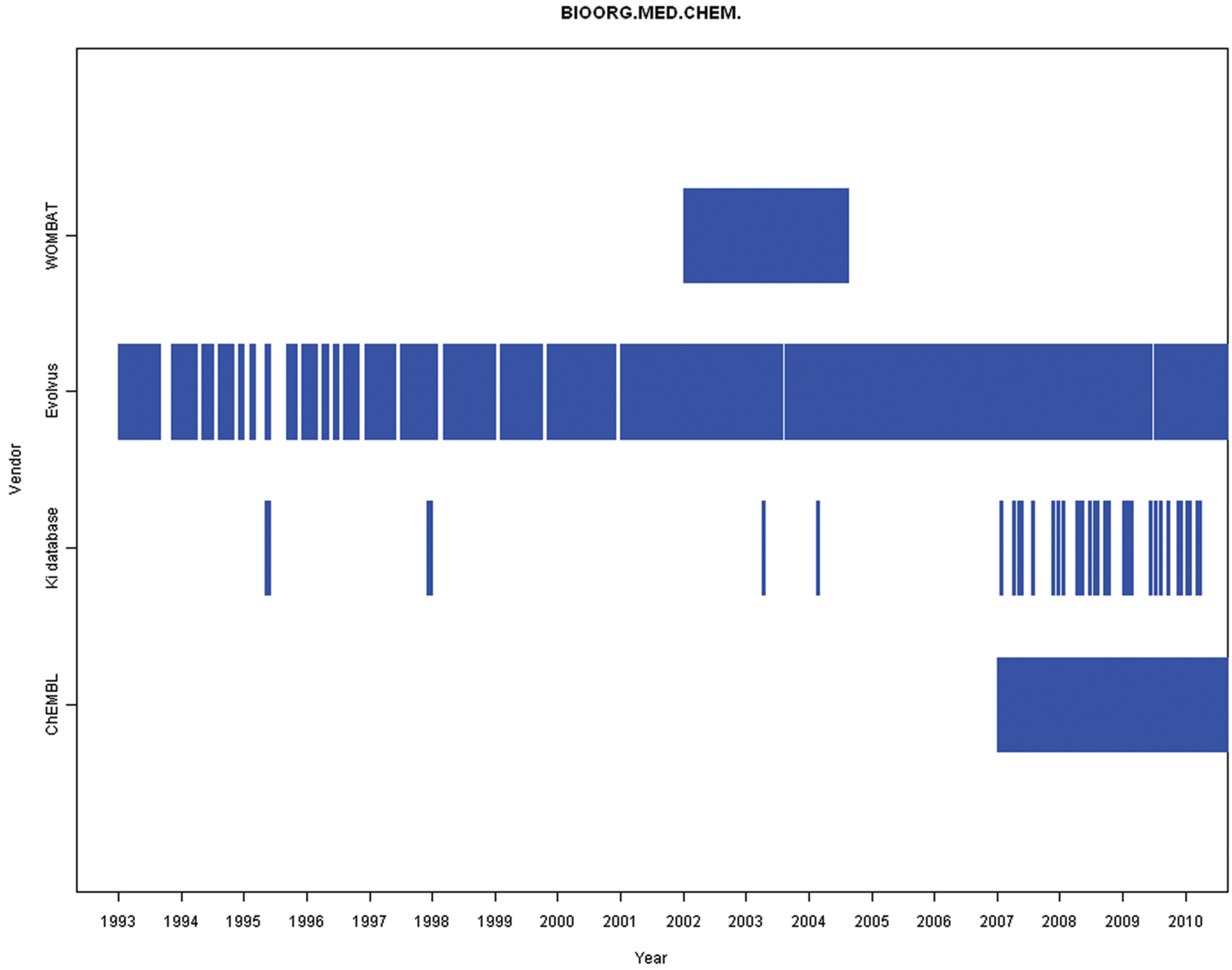


Figure 2. Citation timeline for Bioorganic & Medicinal Chemistry. Each blue bar corresponds to an issue from which at least one article is cited by the given vendor. Similar illustrations for other journals cited by more than one vendor are given in Supporting Information Figure S1.

release 10 published when this article was under preparation improving the target diversity of the latter database.[34] The distribution of activities of the larger databases (ChEMBL, WOMBAT, and Evolvus) in target classes follows the Metabase averages more closely. Table S4 in the Supporting Information has more detailed information on the distribution of activities at different levels of protein family hierarchies.

The relative popularity of target classes (as measured by the amount of activity data published) has not been constant over the past 25 years (Figure 1a). One clear trend visible in Figure 1a is the growth in data for enzymes whose share of activities has risen from around 25−30% in the 1980s to around 40% in the year range 2009−2010. On the other hand, the relative interest in ion channel ligands has dwindled markedly from 20 to 30% of all activities in the 1980s to less than 10% in recent years. Among enzymes, the largest relative growth has been for transferases (Figure 1b) which include protein kinases − a target group of huge research interest especially for their significance in oncology.

Data Discrepancy. 79.3% of the activities extracted from journal articles can only be found in one vendor database. This is despite many of the central medicinal chemistry journals are followed by most of the vendors. Part of the explanation to the relatively low overlap lies in the fact that only restricted time ranges of a particular journal have been extracted by the vendors (Figure 2). Out of the 43,174 distinct articles, 25,194 (58.4%) are cited only by one vendor. In comparison, there are 22,190 articles from which unified activity values had been extracted, and out of these 10,912 (49.2%) are cited by only one vendor.

Since the fraction of activities unique to a vendor is higher than the fraction of articles unique to a vendor, we investigated whether the same data had been extracted by all vendors from an article cited by more than one vendor. There were 11,278 articles with unified activity data that had been cited by more than one vendor. From only 410 (3.6%) of those all vendors had extracted exactly the same activity data. This means that for the remaining 10,868 articles, there has been at least some discrepancy in the data extraction process between the vendors. A total of 898,277 distinct activities have been extracted from the 11,278 articles but only 199,705 (22.2%) of these by all vendors citing the given article. Consequently, the overlap between vendors in terms of their

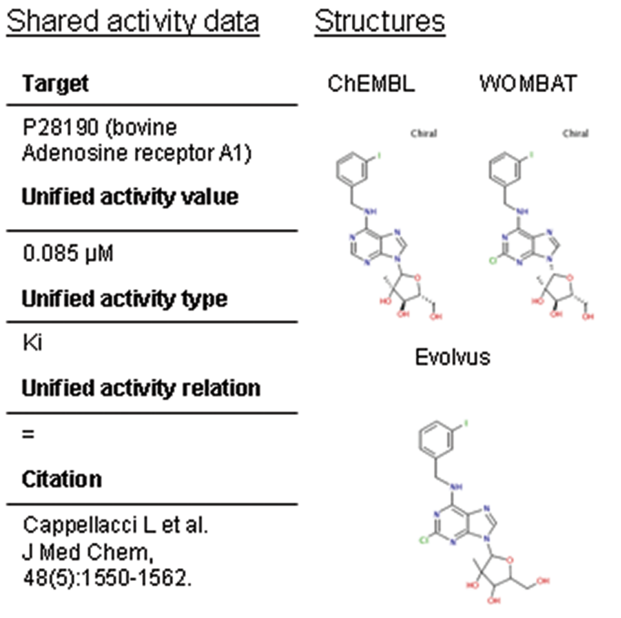


Figure 3. An example of small molecule discrepancy between data vendors. The activity data on the left have been extracted by all three vendors from the same article, but the small molecule structure (compound number three in the original paper) associated with the activity is different with all three vendors. Atom connectivity is correct for WOMBAT and Evolvus, while the chloride atom is missing in the ChEMBL molecule representation. The stereochemistry is incorrect or missing in all three.

activity and small molecule contents would be greater than reported above if all vendors had extracted the data from all articles the same way. In the following, we give examples of data discrepancies on a single activity variable, i.e. cases where activities extracted by different vendors from the same article differ only on the contents of one variable, e.g. target protein.

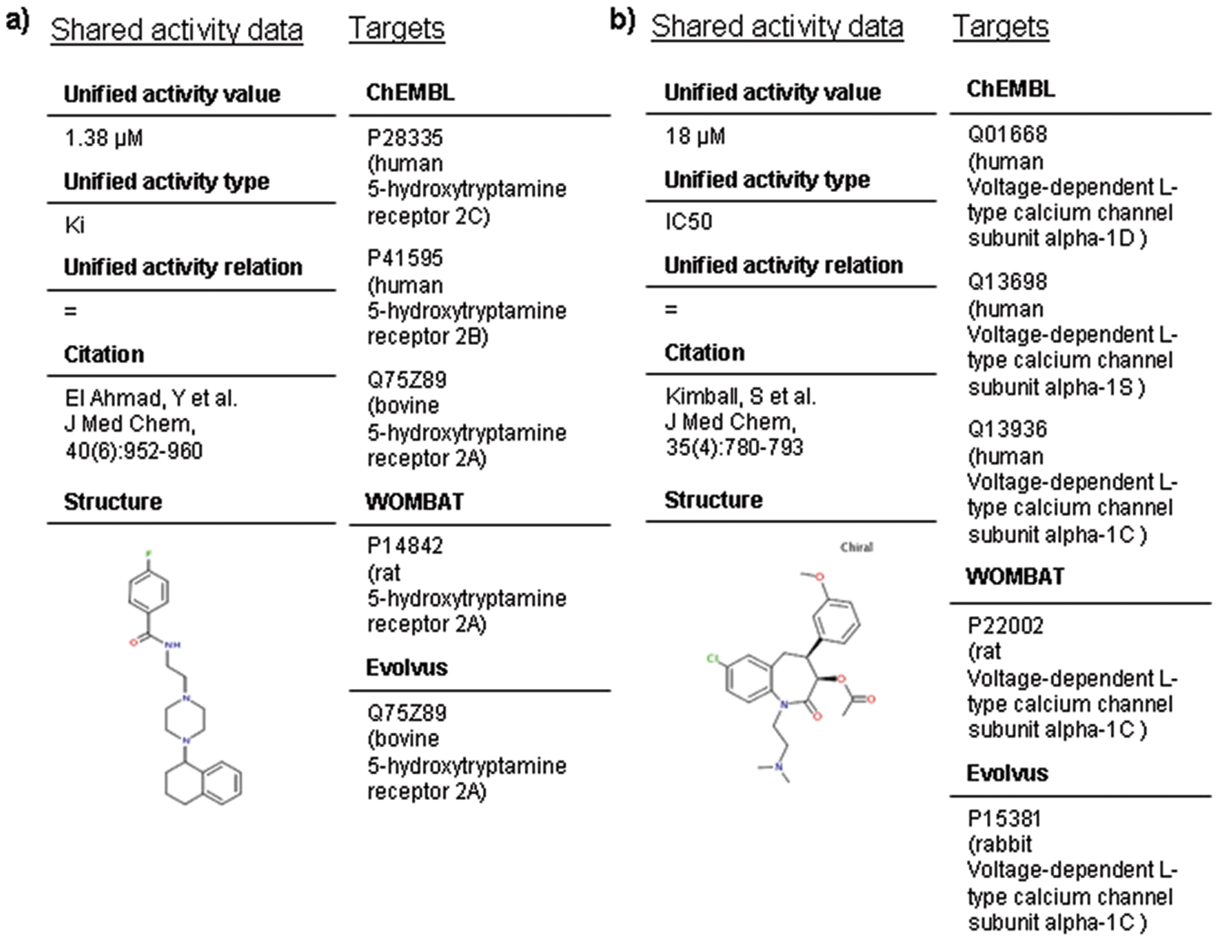
One crucial step when extracting activity data from an article is drawing the small molecule structure correctly. To estimate the degree of discrepancy between vendors, we grouped activities where all vendors citing an article had extracted identical values for all activity fields, but the structure involved with the activity was different. The assumption was that the difference in structure was due to mistakes made by one or more of the vendors when drawing the small molecule. There were 75,278 such groupings corresponding to 180,275 distinct activities (see Figure 3 for an example). For 51,050 (67.8%) of the groupings, the atom connectivity of the structures was the same but a different stereoisomer had been provided by the vendors. Ignoring stereochemistry left us with 24,228 groups corresponding to 57,901 distinct activities where the atom connectivity was different between vendors. Based on this analysis we cannot tell which of the vendor structures are correct, but only highlight cases where the small molecule structure *might* be wrong. The share of activities with a *potentially* wrong molecule structure or stereoisomer (180,275 out of 2,041,996; 8.8%) is high but still a conservative estimate since any mistakes in molecule structures extracted from articles cited by only one vendor could not be analyzed with this approach. Since most of the structure discrepancies are related to stereochemistry, ignoring this would be a partial solution in applications where knowing the exact stereoisomer is not crucial (e.g., tools using most 2D fingerprints).

A similar analysis was made for the protein targets, i.e. Uniprot accessions. Activity data extracted from the same article were merged into groups where the only variable differing between vendors was the Uniprot accession. There are 49,387 such groups corresponding to 123,339 distinct activities. Often the different accessions are for different orthologs of the same protein. This was the case for 32,383 groups, while in the remaining 17,004 groups nonortholog proteins were assigned as targets. The target proteins are often paralogs of each other, for example different serotonin receptor subtypes are given as targets by different vendors for the same activity (Figure 4a). In this particular case, the assay used to measure the activity was not receptor subtype specific, and it remains unclear which of the subtypes were actually targeted by the ligand. This can lead to problems, e.g. when one wants to use the ligand in Figure 4a as a reference molecule in ligand-based virtual screening for new Serotonin receptor 2A inhibitors. Since there is no certainty that the ligand binds the said subtype, the results could be misleading. Another example of target discrepancy is given in Figure 4b where the target is an ion channel composed of several individual protein subunits. ChEMBL assigns each subunit as a target, whereas WOMBAT and Evolvus cite just one apparently arbitrarily chosen subunit. In light of the latter example, when searching protein complex ligands one should search with all complex subunits in order not to miss any.

ChEMBL contains a target confidence field telling the level of detail known about the actual binding partner. Data from assays measuring activity against a single protein are given a value of nine, whereas the activities in Figure 4a have a confidence scores ranging from four to five meaning it is not known which of the homologues the ligand actually binds. Filtering data out by the confidence score would therefore be helpful if higher confidence is needed. Unfortunately the other vendor databases do not comprise a similar field.

As stated above, discrepancies and mistakes in data extraction mean that the figures for activities and molecules given in Tables 1 and 2 are inflated. By considering stereoisomers of the same molecule as identical (i.e., ignoring stereochemistry) and identifying protein targets by their recommended protein names instead of by their Uniprot accessions (i.e., considering orthologous proteins from different species as the same target) we should get a more accurate figure for the frequencies and uniqueness of both activities and molecules. This leads the total number of distinct activities to drop by 7.8% (from 2,041,996 to 1,882,337) and the share of activities unique to a single vendor to drop from 87.3% to 82.4%. Correspondingly, ignoring stereochemistry leads to an 8.7% drop in the number of molecules, while the share of small molecules unique to a single vendor drops from 88.9% to 86.3%.

It should be noted that the molecule and activity counts reported in the current work are a snapshot of the situation when work on this article began. Regular updates and growth of the vendor databases lead the exact figures to change. We argue though that the more general findings reported here (discrepancies and relatively low overlap of the vendor databases) are going to hold true for several update cycles to come.



a) Shared activity data

Unified activity value

1.38 µM

Unified activity type

Ki

Unified activity relation

=

Citation

El Ahmad, Y et al.
J Med Chem,
40(6):952-960

Structure

Targets

ChEMBL

P28335 (human 5-hydroxytryptamine receptor 2C)

P41595 (human 5-hydroxytryptamine receptor 2B)

Q75Z89 (bovine 5-hydroxytryptamine receptor 2A)

WOMBAT

P14842 (rat 5-hydroxytryptamine receptor 2A)

Evolvus

Q75Z89 (bovine 5-hydroxytryptamine receptor 2A)

b) Shared activity data

Unified activity value

18 µM

Unified activity type

IC50

Unified activity relation

=

Citation

Kimball, S et al.
J Med Chem,
35(4):780-793

Structure

Targets

ChEMBL

Q01668 (human Voltage-dependent L-type calcium channel subunit alpha-1D)

Q13698 (human Voltage-dependent L-type calcium channel subunit alpha-1S)

Q13936 (human Voltage-dependent L-type calcium channel subunit alpha-1C)

WOMBAT

P22002 (rat Voltage-dependent L-type calcium channel subunit alpha-1C)

Evolvus

P15381 (rabbit Voltage-dependent L-type calcium channel subunit alpha-1C)

Figure 4. Examples of vendors assigning different protein targets for the same activity. a) The activity data in this example were measured with a subtype inspecific bovine 5-HT2 receptor assay. In this particular case reporting all subtypes as targets (ChEMBL) — and not picking one arbitrarily — should be considered the correct approach although the species is incorrect for two of the ChEMBL targets. b) An example of an activity where the target is a protein complex (an ion channel) and not an individual protein. ChEMBL has assigned each individual channel subunit as a target, whereas WOMBAT and Evolvus cite only one. Again, the ChEMBL approach should be considered the correct one in contrast to picking one subunit arbitrarily.

CONCLUSIONS

Combining data from several bioactivity databases clearly has its advantages since each of them contains unique data. Despite the increase in public domain data during the past few years, the commercial databases still offer complementary data. Since the data sources used by the vendors differ, the lack of complete overlap is hardly surprising. However, the vendors have discrepancies with data from the same articles. These lead to a smaller overlap than it would be if the vendors had been able to reduce such data extraction faults. One step toward higher data quality could be to cross-check the data with other sources in a similar way we have done in the current work, identify journal articles with discrepancies, and correct any mistakes in molecule structures or any other activity parameters by recurating the data. We note that this approach works only where an activity is from a data source cocited by two or more vendors and does not identify errors where the data source is cited by one vendor only.

Care should be taken when using the databases to pick molecules active against a specific target. The target information is sometimes on the protein (sub)family level, and there is no guarantee that the specific protein given in the activity information is in reality bound by the small molecule. ChEMBL has addressed the problem by assigning a target confidence score to its activities. To us this is a very helpful approach which could be adopted more widely.

The amount of data accumulated in these databases is enormous and invaluable to the drug discovery effort. In addition to collecting more data, also the extraction process may be further improved to increase the confidence in the data. Having two or more curators extract data from a given article independently and cross-checking the results would be a very effective yet costly strategy to decrease the error rate in the extraction step. However, we do not know whether this is an already established procedure at data vendors. We would still like to note that also mistakes in the cited articles and patents themselves decrease data quality, and there is little the data vendors can do about this.[13]

Fortunately the scientific community has recently started to pay more attention to the data quality issue,[35] and initiatives such as MIABE[36] are showing the way for industry-wide standards for bioactivity data storage and reporting. Our hope is that the findings presented in the current work will in their part help raise awareness on the data quality in these databases and even improve the practices in collection of bioactivity data.

■ ASSOCIATED CONTENT

Supporting Information. List of PubChem bioassays used in the work (Table S1), effect of structure standardization on molecule entity counts (Table S2), overlap of vendor databases (Table S3), number of activities at different levels of protein family hierarchy (Table S4), and citation timelines for journals cited by more than one vendor (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49 69 15 03 8158. Fax: +49 69 15 03 188. E-mail: pekka.tiikkainen@merz.de.

■ ACKNOWLEDGMENT

Dr. Yusuf Tanrikulu is credited for reading through the manuscript and giving valuable comments.

■ REFERENCES

(1) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.

(2) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.

(3) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* [Online early access]. DOI: 10.1093/nar/gkr777. Published Online: Sept 23, **2011**.

(4) European Bioinformatics Institute. ChEMBL. https://www.ebi.ac.uk/chembldb/index.php (accessed February 7, 2011).

(5) Bender, A. Databases: Compound bioactivities go public. *Nat. Chem. Biol.* **2010**, *6*, 309–309.

(6) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805–815.

(7) Southan, C.; Varkonyi, P.; Muresan, S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminf.* [Online] **2009**, *1*, Article 10. http://www.jcheminf.com/content/1/1/10 (accessed Feb 24, 2011).

(8) Taboureau, O.; Nielsen, S. K.; Audouze, K.; Weinhold, N.; Edsgard, D.; Roque, F. S.; Kouskoumvekaki, I.; Bora, A.; Curpan, R.; Jensen, T. S.; Brunak, S.; Oprea, T. I. ChemProt: a disease chemical biology database. *Nucleic Acids Res.* **2010**, *39*, D367–D372.

(9) Technical University of Denmark. Chemprot. http://www.cbs.dtu.dk/services/ChemProt/ (accessed May 10, 2011).

(10) Sunset Molecular. WOMBAT database. http://www.sunset-molecular.com/ (accessed April 13, 2011).

(11) Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B. A.; Suzek, T. O.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. H. An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **2010**, *38*, D255–D266.

(12) National Center for Biotechnology Information. The PubChem Project. http://pubchem.ncbi.nlm.nih.gov/ (accessed October 18, 2010).

(13) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*, 1st ed.; Oprea, T. I., Ed.; Wiley: New York, NY, 2005; Vol. 23, pp 223−239.

(14) *Pipeline Pilot*, version 8.0; Accelrys, Inc.: San Diego, CA, 2011.

(15) Evolvus. Discovery Informatics. http://www.evolvus.com/di.htm (accessed April 21, 2011).

(16) Roth, B. L.; Lopez, E.; Patel, S.; Kroeze, W. K. The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches? *The Neuroscientist* **2000**, *6*, 252–262.

(17) NIMH Psychoactive Drug Screening Program. Ki Database. http://pdsp.med.unc.edu/kidb.php (accessed Mar 14, 2011).

(18) University of Alberta. Drugbank. http://www.drugbank.ca/index.html (accessed Nov 17, 2010).

(19) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

(20) BindingDB information page. http://www.bindingdb.org/bind/info.jsp (accessed Sept 19, 2011).

(21) International Union of Pure and Applied Chemistry. The IUPAC International Chemical Identifier, version 1.03. http://www.iupac.org/inchi (accessed May 19, 2011).

(22) Uniprot Consortium. The Universal Protein Resource. http://www.uniprot.org (accessed March 18, 2011).

(23) The National Center for Biotechnology Information. UniGene: An Organized View of the Transcriptome. http://www.ncbi.nlm.nih.gov/unigene (accessed July 14, 2011).

(24) The National Center for Biotechnology Information. Sequence Identifiers: A Historical Note. http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html (accessed July 6, 2011).

(25) The Universal Protein Resource. Uniprot Identifier Mapping File. ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz (accessed January 12, 2011).

(26) The Universal Protein Resource . Protein naming guidelines. http://www.uniprot.org/docs/nameprot (accessed Sept 28, 2011).

(27) IUPHAR Committee on Receptor Nomenclature and Drug Classification. IUPHAR database of receptors and ion channels. http://www.iuphar-db.org/ (accessed Apr 26, 2011).

(28) Harmar, A. J.; Hills, R. A.; Rosser, E. M.; Jones, M.; Buneman, O. P.; Dunbar, D. R.; Greenhill, S. D.; Hale, V. A.; Sharman, J. L.; Bonner, T. I.; Catterall, W. A.; Davenport, A. P.; Delagrange, P.; Dollery, C. T.; Foord, S. M.; Gutman, G. A.; Laudet, V.; Neubig, R. R.; Ohlstein, E. H.; Olsen, R. W.; Peters, J.; Pin, J. P.; Ruffolo, R. R.; Searls, D. B.; Wright, M. W.; Spedding, M. IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.* **2009**, *37*, D680–D685.

(29) Swiss Institute of Bioinformatics. ExPASy - ENZYME. http://expasy.org/enzyme/ (accessed May 3, 2011).

(30) University of California: San Diego. Transporter Classification Database. http://www.tcdb.org (accessed May 11, 2011).

(31) Saier, M. H., Jr.; Yen, M. R.; Noto, K.; Tamang, D. G.; Elkan, C. The Transporter Classification Database: recent advances. *Nucleic Acids Res.* **2009**, *37*, D274–D278.

(32) US National Library of Medicine. Pubmed Home Page. http://www.ncbi.nlm.nih.gov/pubmed/ (accessed Sept 28, 2011).

(33) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5*, 993–996.

(34) The ChEMBL database blog. http://chembl.blogspot.com/2011/09/integration-of-filtered-set-of-pubchem.html (accessed Sept 21, 2011).

(35) Williams, A. J.; Ekins, S. A quality alert and call for improved curation of public chemistry databases. *Drug Discovery Today* **2011**, *16*, 747–750.

(36) Orchard, S.; Al-Lazikani, B.; Bryant, S.; Clark, D.; Calder, E.; Dix, I.; Engkvist, O.; Forster, M.; Gaulton, A.; Gilson, M.; Glen, R.; Grigorov, M.; Hammond-Kosack, K.; Harland, L.; Hopkins, A.; Larminie, C.; Lynch, N.; Mann, R. K.; Murray-Rust, P.; Lo Piparo, E.; Southan, C.; Steinbeck, C.; Wishart, D.; Hermjakob, H.; Overington, J.; Thornton, J. Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discovery* **2011**, *10*, 661–669.