

Analysis of contents in proprietary and public bioactivity databases



Pekka Tiikkainen and Lutz Franke

Repurposing, Drug Discovery and Development, Merz Pharmaceuticals GmbH, Frankfurt am Main, Germany

Abstract

Bioactivity databases are essential in drug discovery. Both public and commercial databases are available with up to millions of data points each.

Objectives of our work were

1. to combine all data sources (vendors) available to us into a single resource to be used in building drug target models.
2. to determine the overlap of commercial and public vendors to justify the investment into the former
3. to identify range of inconsistency between vendors citing the same article

The main findings were

1. Commercial vendors include data not found in public sources and *vice versa*.
2. Despite vendors share several journals as sources, the range of volumes and issues covered varies.
3. Data extracted by different vendors from the same article is often inconsistent.

In conclusion, using databases by different vendors is still justifiable since the data overlap is not complete. It should be noted that this can be partially explained by the inconsistencies and errors we have found in the source data.

Overview

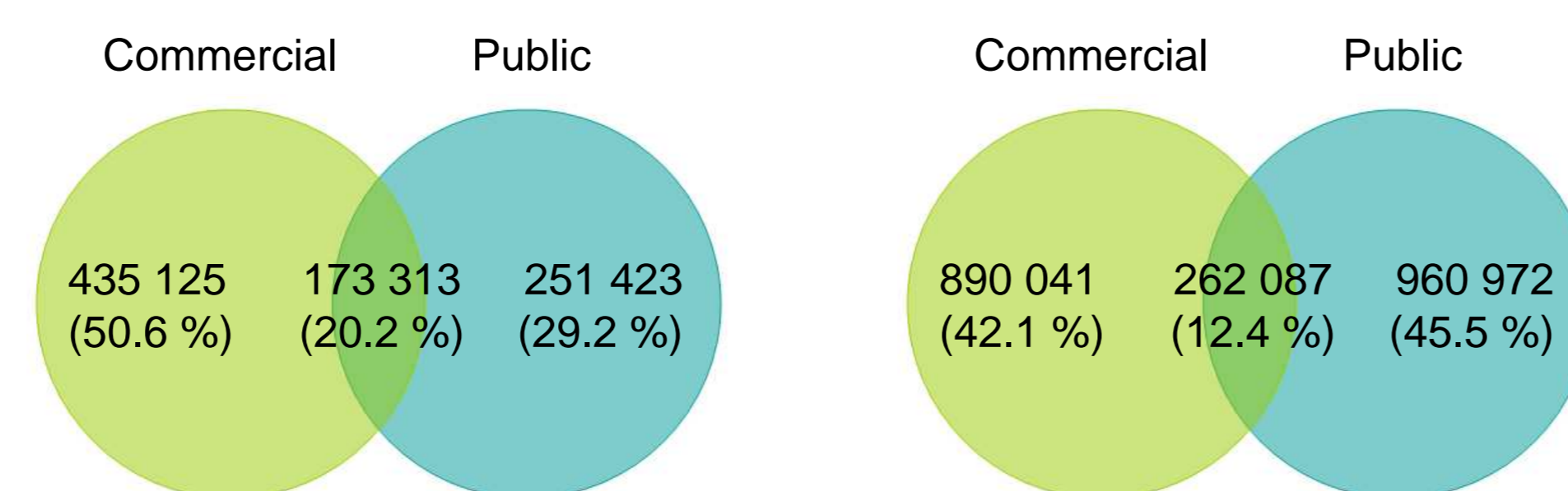
General statistics

Vendor	Total datapoints	Unique molecules	Targets (Uniprot IDs)	Molecules associated with activity**	Activities**
BindingDB, v. 26.11.10 (P)	603 282	260 970	3 070	228 987	492 378
ChEMBL, build_09 (P)	2 914 811	585 225	4 554	306 626	904 841
Drugbank, version 2.5 (P)	10 738	3 869	4 216	N/A ¹	N/A ¹
Liceptor, v. 2010 (C)	9 234 819	1 899 413	1 715	469 166	837 628
Ki database, v. 21_01_11 (P)	47 821	3 887	482	3 381	30 739
PubChem (P)	201 797	89 993	298	89 978	200 889
WOMBAT, 2011.01 (C)	649 547	251 240	2 677	186 732	378 743
Total (% unique to single vendor)	13 662 815	2 493 521 (85.8 %)	9 166 (60.3 %)	859 861 (69.8 %)	2 113 100 (75.8 %)

P = public, C = commercial.

¹ Drugbank does not contain quantitative bioactivity data.

Molecule overlap* Overlap of activities**



* Numbers are for molecules associated with an activity.

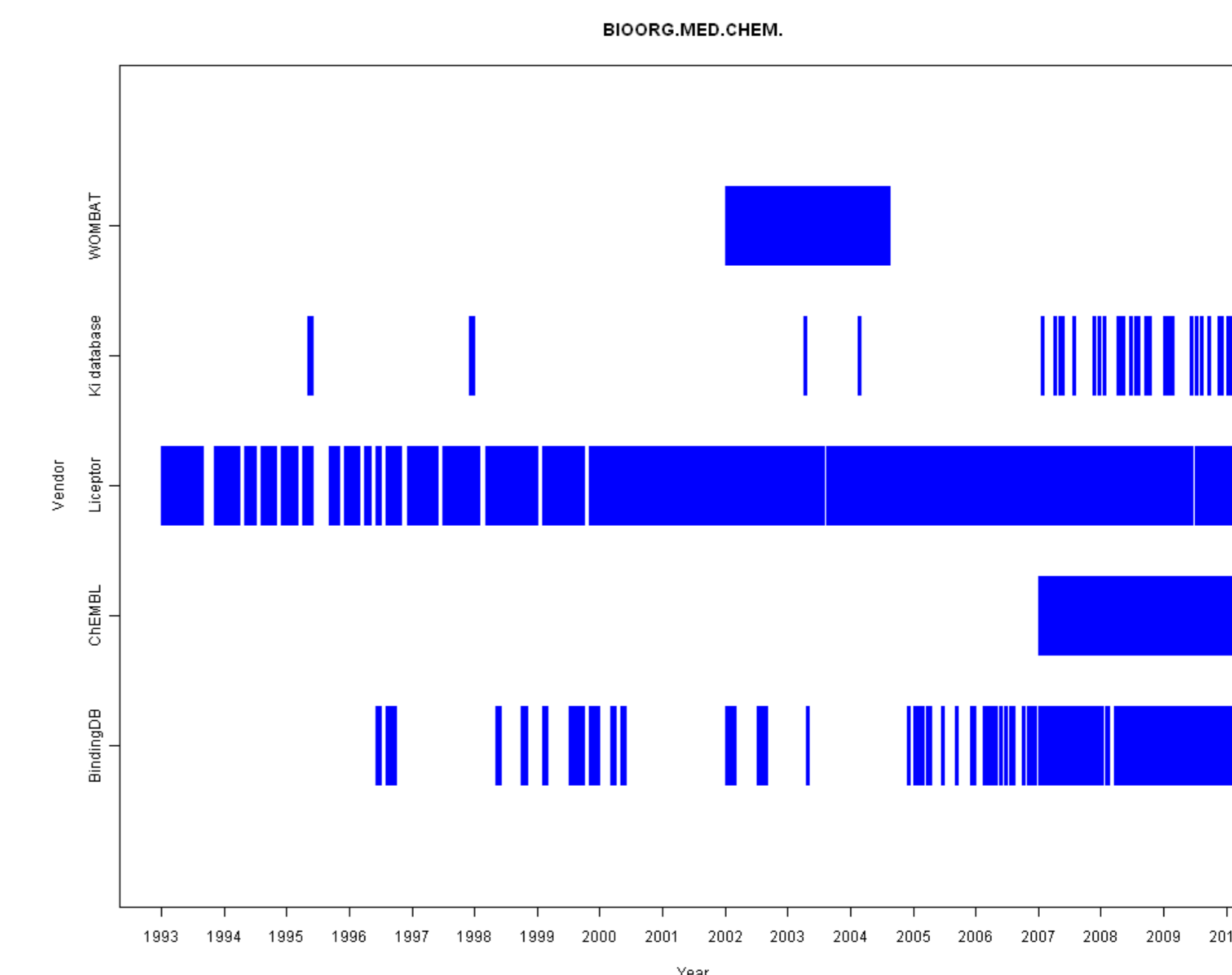
** An activity is defined as a unique combination of Uniprot ID, small molecule, activity value, activity type and activity relation.

Article timeline

Data vendors differ largely in the range of issues and journals cited. This – together with the fact that from our datasets Liceptor is the only one with patent data - is the main reason for the low data overlap.

The plot on the right shows Bioorganic & Medicinal Chemistry issues where vendors have cited at least one article.

Although all the five vendors cite the journal, the range of issues and articles covered varies.

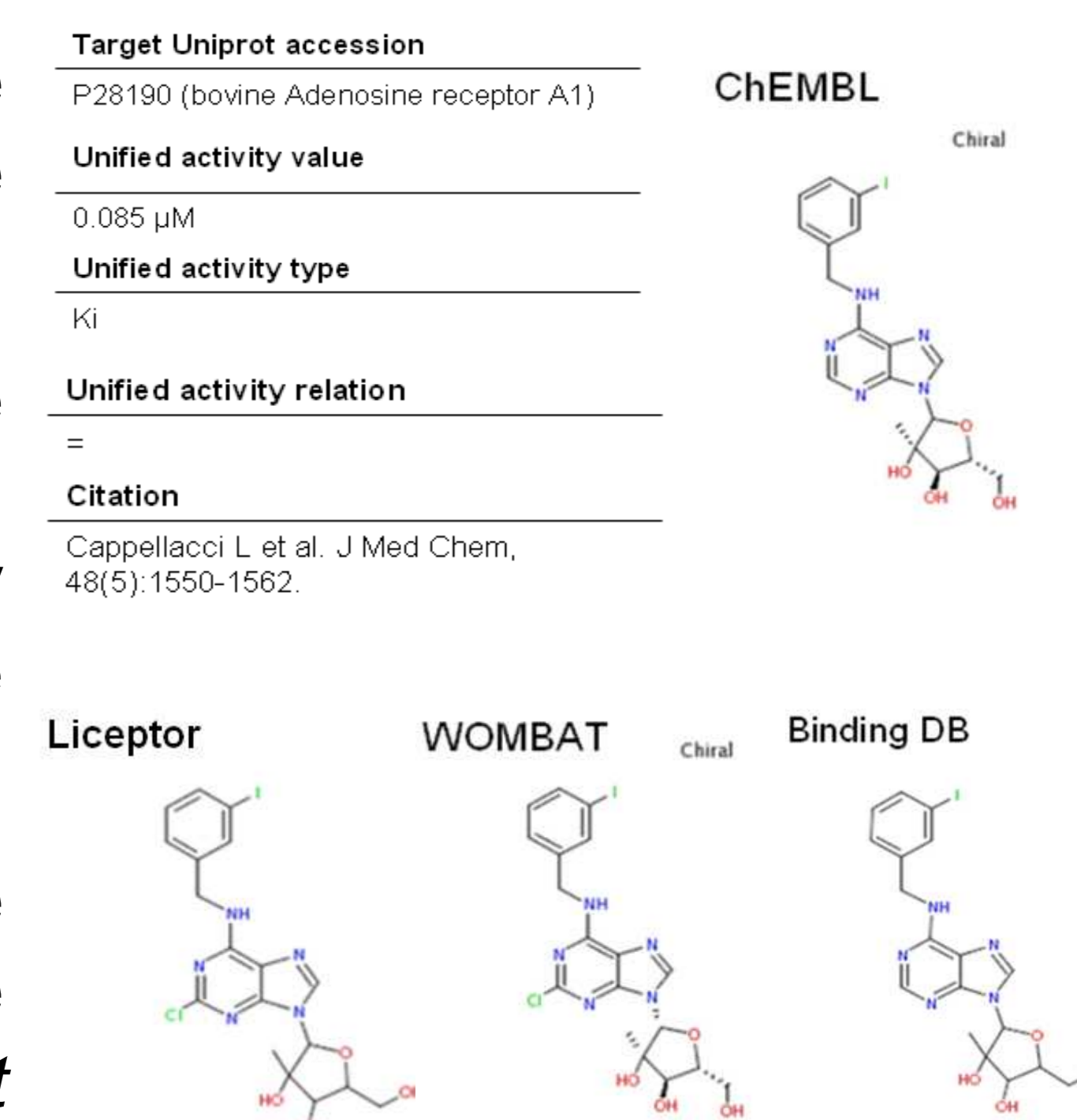


Discrepancies

Another reason for the low overlap are discrepancies in activity data two or more vendors have extracted from the same article.

There are 15 550 articles cited by more than one vendor but exactly the same activity data had been extracted from only 1 898 (12.2 %) of these by all vendors that cite the article.

The example on the right shows a case where four vendors cite the same article and have extracted exactly the same activity data, *except* they have drawn the small molecule differently.



Vendor target preferences

The heatmap shows the relative activity frequencies against two important drug target classes (ion channels and GPCRs). Target classification is based on the IUPHAR website (www.iuphar-db.org).

