Estimating Error Rates in Bioactivity Databases

Pekka Tiikkainen

6th Joint Sheffield Conference on Chemoinformatics

July 24, 2013

Presentation overview

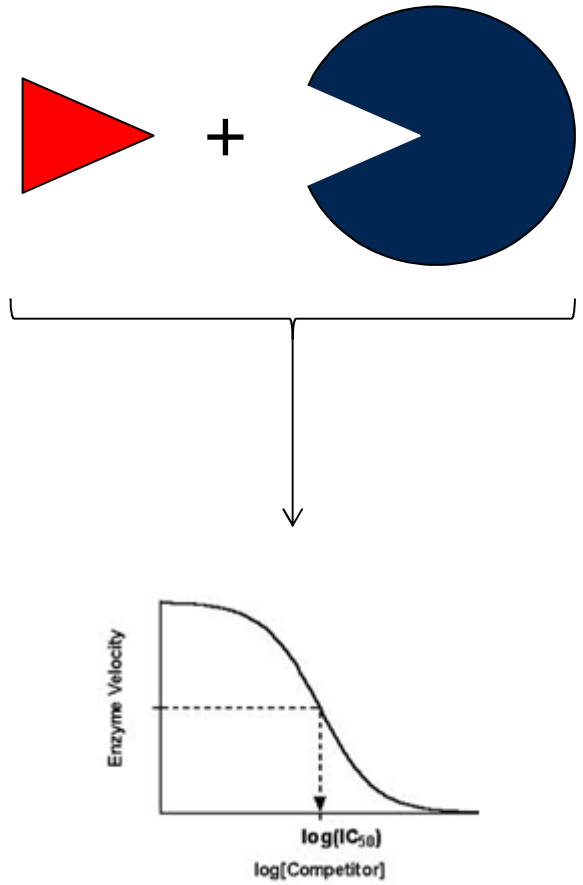
- **Merz Virtual Bioactivity Database**
- **Discrepancies and error rate estimates**
- **Conclusions and summary**

Presentation overview

- **Merz Virtual Bioactivity Database**
- **Discrepancies and error rate estimates**
- **Conclusions and summary**

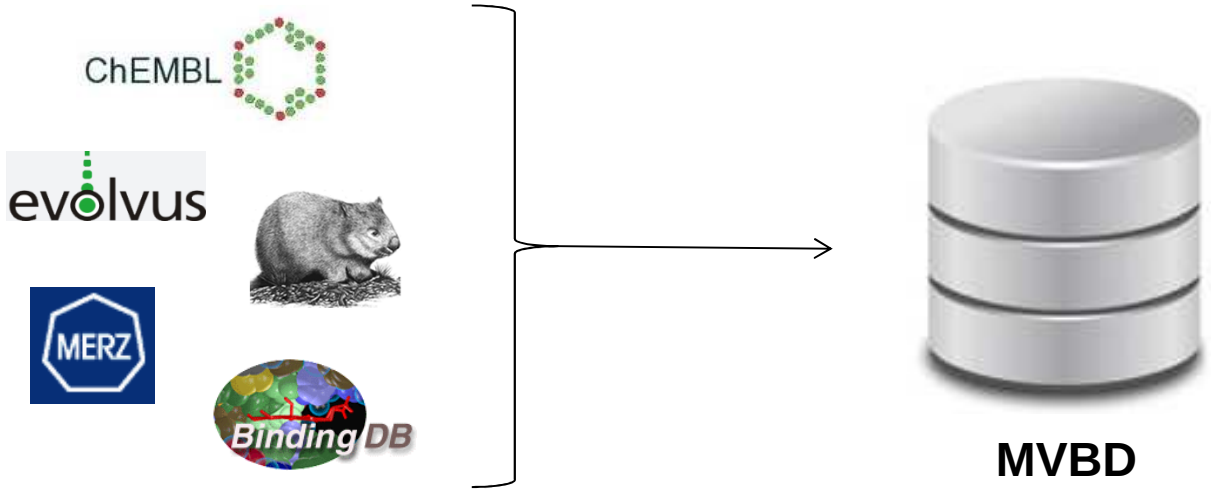
Bioactivity databases

- Central to modern drug discovery.
- Built with a largely manual curation of scientific articles and patents. Some suppliers include screening data sets etc.
- Both commercial and public databases available.
- Pharmacokinetic and pharmacodynamic data.
- In this talk, I will concentrate on activity data where the following parameters have been defined:
  - ligand structure
  - target protein (Uniprot accession)
  - quantitative activity value
  - activity type (Ki, IC50, EC50 etc.)



Merz Virtual Bioactivity Database (MVBD)

- Our central resource for bioactivity data.
- Used for target prediction and enriching other data resources.
- Integrated from public, commercial and in-house data resources.
- Implemented as a MySQL database.



Bioactivity breakdown by target class

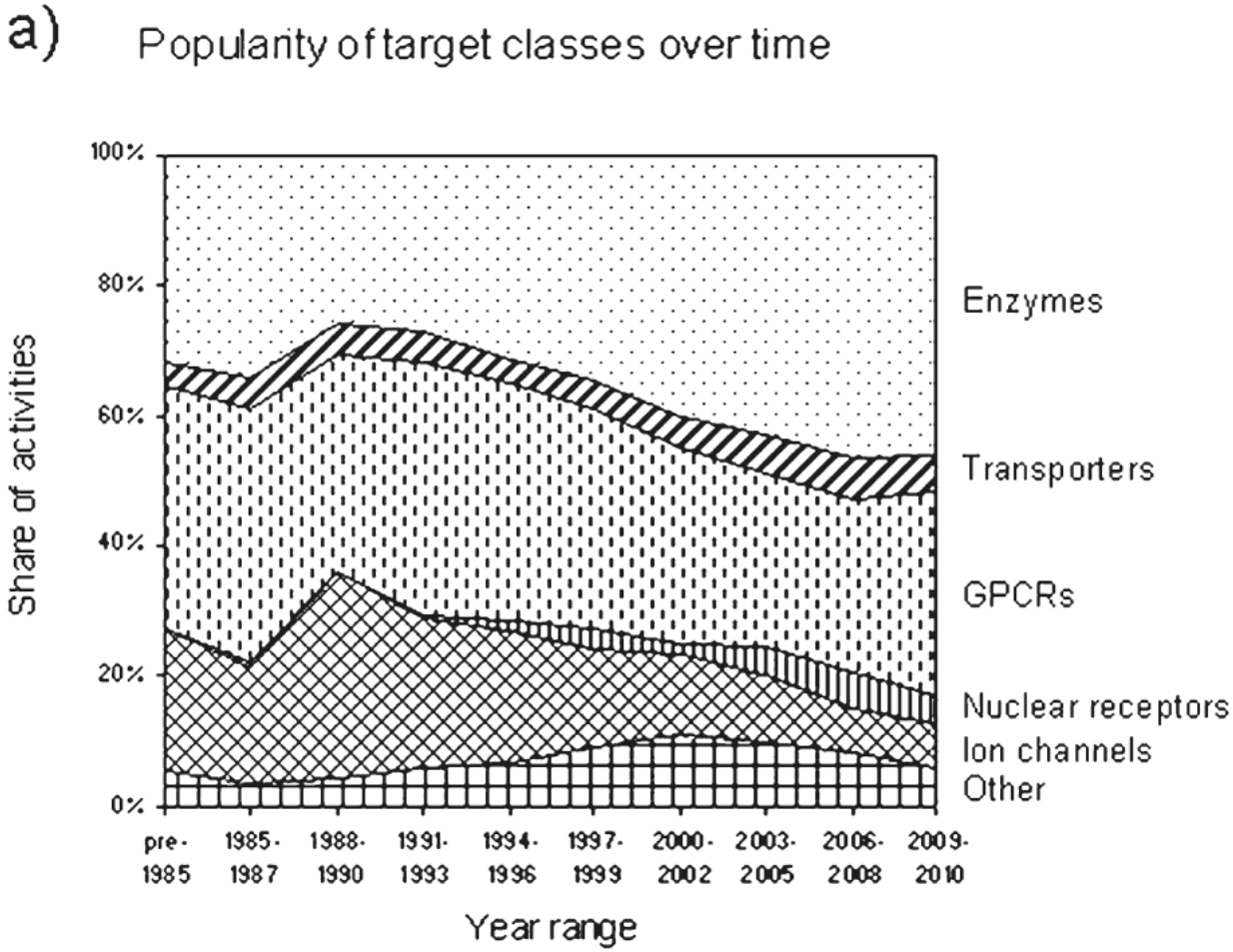
Table 3. Distribution of Activities Across Five Major Protein Classes[a]

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%)

Whereas GPCRs (still) are the largest target class for launched drugs, enzymes are the largest target class in the MVBD.

Tiikkainen and Franke. Analysis of commercial and public bioactivity databases. J Chem Inf Model. 2012 Feb 27;52(2):319-26.

Historical breakdown of target data

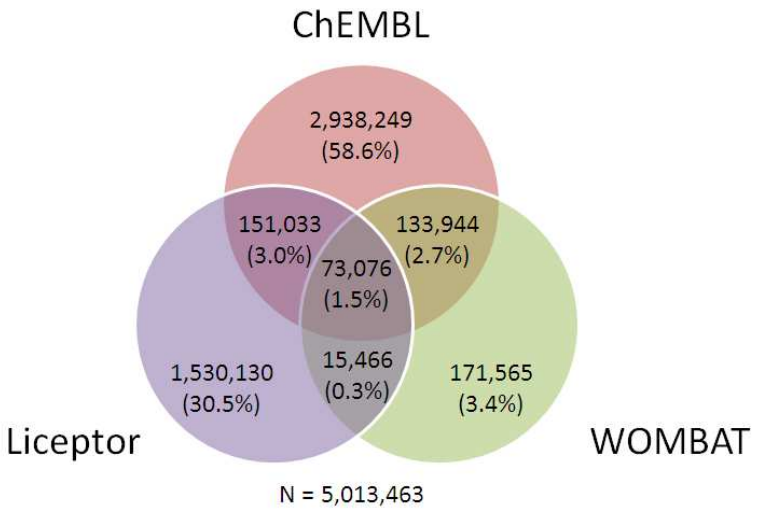


Tiikkainen and Franke. Analysis of commercial and public bioactivity databases. J Chem Inf Model. 2012 Feb 27;52(2):319-26.

Why integrate?

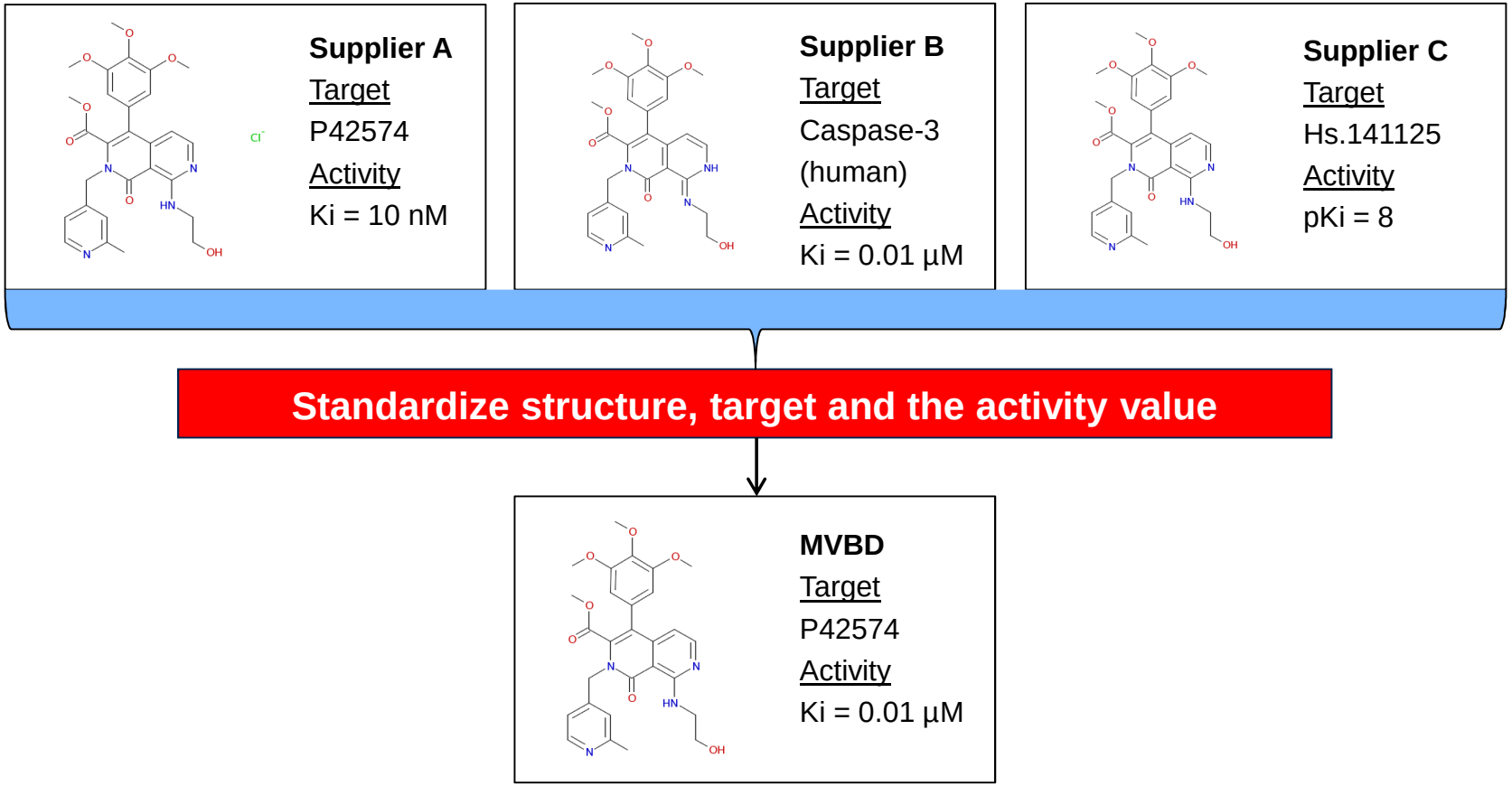
- Much of the bioactivity data is unique to a single database.
- All databases use scientific papers as data sources, but...
  - not all databases cite the same journals
  - the extent a journal is cited varies across databases
  - databases use additional data sources, e.g. Pubchem and screening data sets for ChEMBL, patents for Liceptor

Overlap of bioactivities



Bioactivity standardization

When your data comes from heterogenous sources, standardization becomes extremely important.



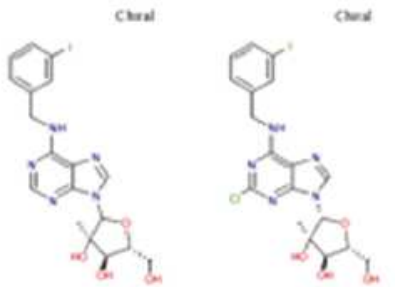
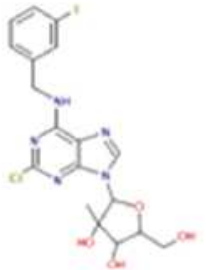
Presentation overview

- Merz Virtual Bioactivity Database
- **Discrepancies and error rate estimates**
- Conclusions and summary

Discrepancies

Standardization and integration allows us to compare the different database suppliers.

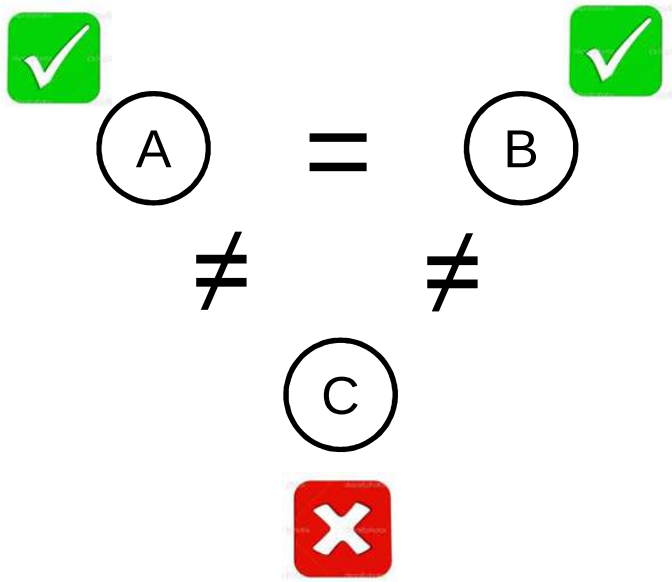
When comparing bioactivities different suppliers have curated from the same article, it is not uncommon to find discrepancies.

Shared activity data	Structures
Target	ChEMBL WOMBAT
P28190 (bovine Adenosine receptor A1)	
Unified activity value	
0.085 µM	
Unified activity type	
Ki	Evolvus
Unified activity relation	
=	
Citation	
Cappellacci L et al. J Med Chem, 48(5):1550-1562.	

Tiikkainen and Franke. Analysis of commercial and public bioactivity databases. J Chem Inf Model. 2012 Feb 27;52(2):319-26.

From discrepancies to error rate estimates

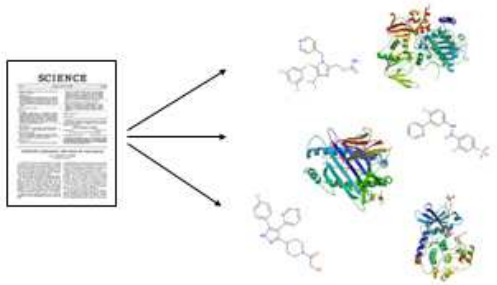
- Discrepancies alone do not tell which of the suppliers is correct.
- However, we can calculate error rate estimates using a special subset of discrepancies: if two database supplier agree on a parameter value while a third one disagrees, the latter value is considered incorrect.
- Underlying assumption is that all suppliers have independently curated the articles.



Error rate estimation workflow

1 2,184 articles cited by all three database vendors.

2 For each article, activity records extracted by each vendor were listed.



3 An article's activity records were grouped using one of the activity parameters as pivot variable at a time. Groups where at least one vendor had more than one pivot value were ignored.

Target	Activity value	Activity type	Molecular structure (pivot variable)		
			Liceptor	WOMBAT	ChEMBL
P11712	17 nM	Ki	Cmpd X	Cmpd X	Cmpd X
S45829	5 µM	IC50	Cmpd Z	Cmpd Z	Cmpd Y
R24582	1.7 nM	Kd	Cmpd B	Cmpd C	Cmpd Z

Shared activity parameters

4 By comparing the pivot values, discrepant activities were identified.

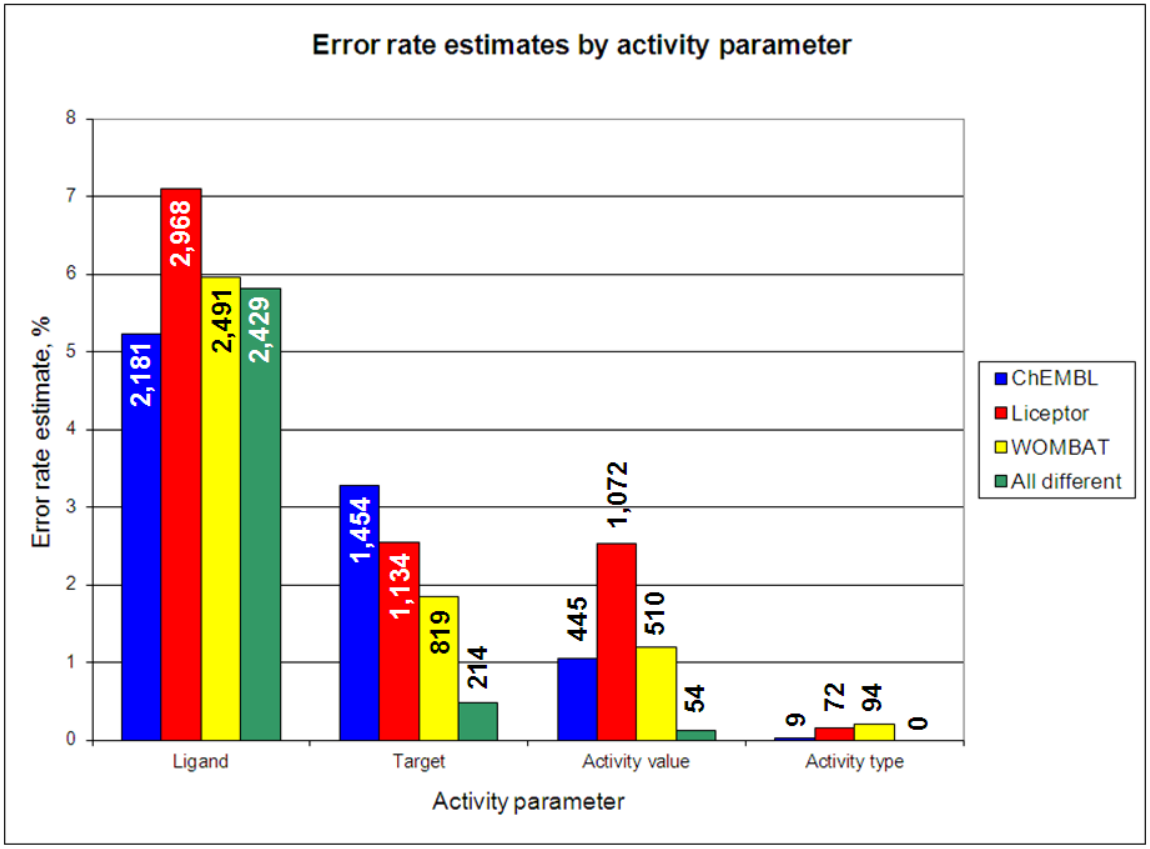
All pivot values agree -> all values assumed to be **correct**.

Two of the vendor pivot values are identical while the third one not
-> the third vendor's value is assumed to be **incorrect**.

All three pivot values are different
-> at most one value can be correct but one can't say which one.

Error rate estimates

Calculating discrepancy frequencies gives us error rate estimates.



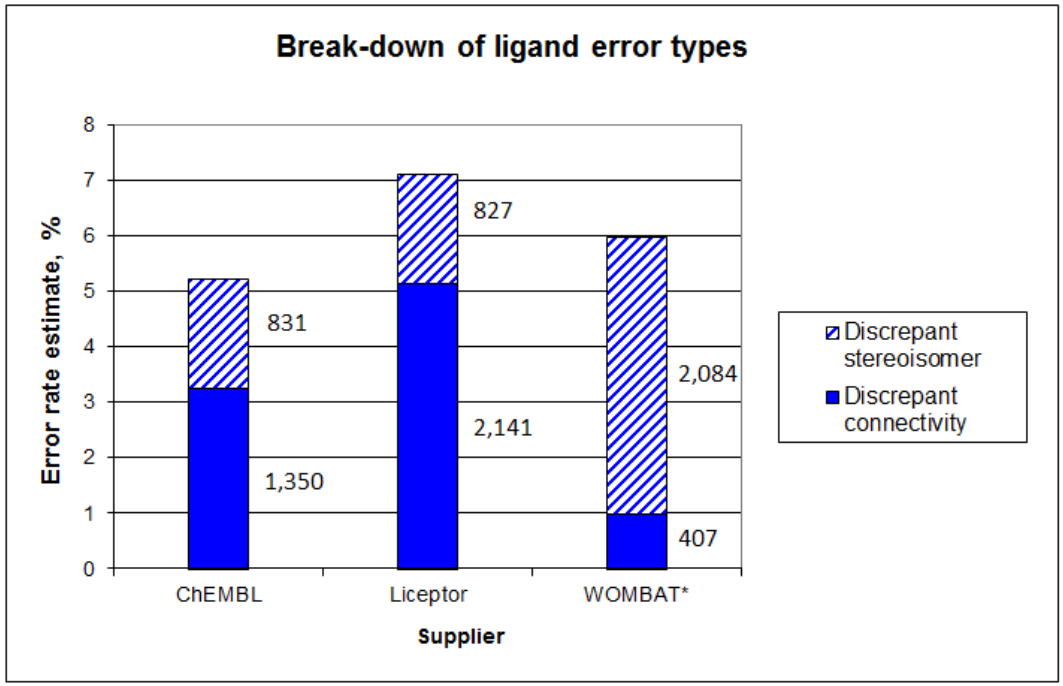
Figures inside and above the bars indicate the absolute number of bioactivities.

Types of ligand errors

Discrepancies in ligand structures can be split into two categories:

1) discrepancies in atom connectivity and

2) atom connectivity identical but discrepant stereochemistry



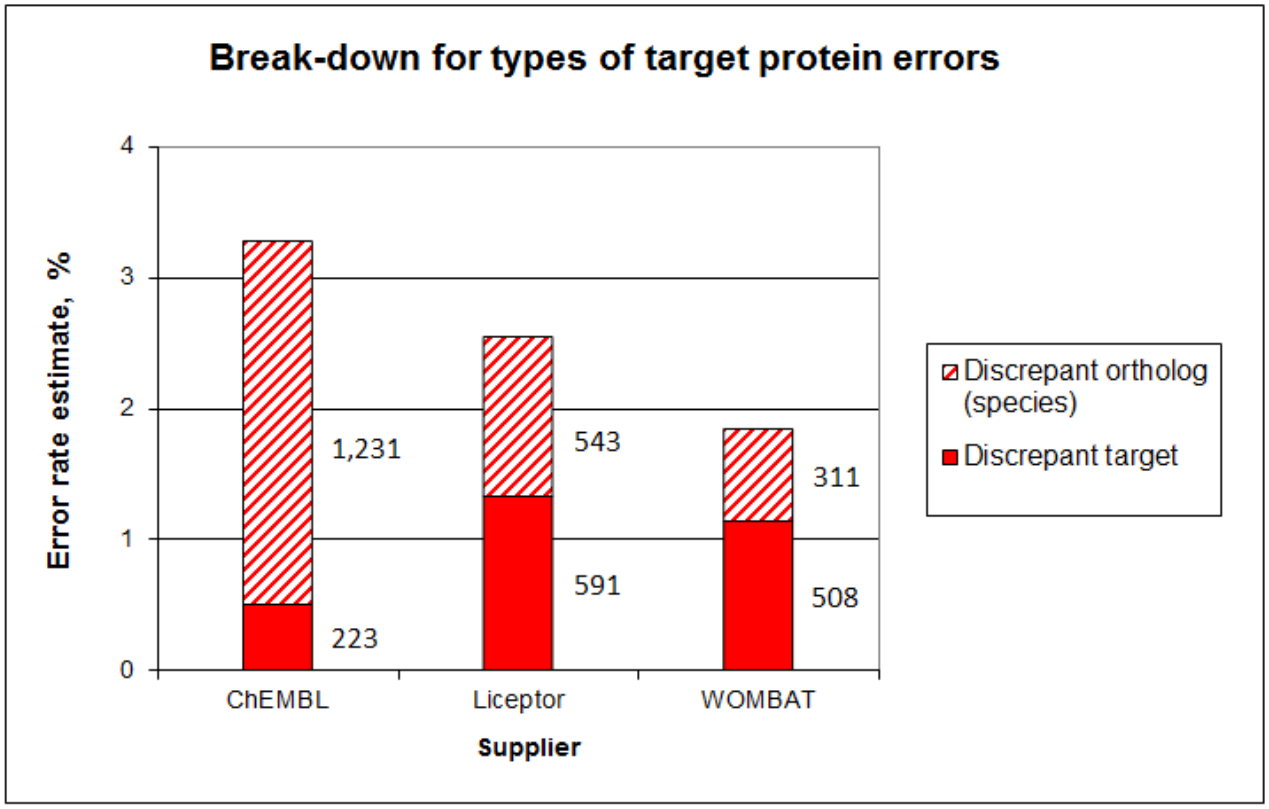
* Majority of stereochemistry discrepancies in WOMBAT is probably due to lack of any stereochemical features in other databases.

Types of target errors

Also target discrepancies can be split into two categories:

1) target protein itself is discrepant (e.g. 5-HT1a vs. 5-HT2a)

2) target protein is correct but discrepant ortholog (source species)



Validating the approach, part 1

For the credibility of the approach, it was necessary to test how often the underlying assumption holds.

For each activity parameter (excl. activity type) and supplier, we picked five activities where the supplier had provided a discrepant value (i.e. 45 activities).

These activities were manually checked from the original source articles.

In 37 cases (**82.2%**), the discrepant activity value turned out in fact to be wrong -> **assumption correct**.
In 3 cases (6.7%), the opposite was true, and the discrepant supplier in fact had the correct value -> **assumption incorrect**.
In 5 cases (11.1%), we could not draw a conclusion since the source article was lacked clarity on the exact parameter value.

Validating the approach, part 2

A more extensive validation was performed by the ChEMBL team while checking discrepancies identified in ChEMBL release 14.

Parameter	Set	Results
Ligand structure	1,936 ligands (corresponding to 2,181 activities) discrepant only in ChEMBL.	310 (16.0%) correctly curated in ChEMBL while the remaining 1,626 (84.0%) required some changes.
	1,486 ligands (2,429 activites) where all suppliers disagreed.	280 ligands (18.8%) correctly curated in ChEMBL. For 1,206 ligands (81.2%) some changes had to be made.
Activity type/value	259 cases checked so far.	In 83 cases (32.0%), ChEMBL had the correct activity value and type. For 68.0% of the cases, some corrections were made.
Target	764 bioactivities where ChEMBL was the sole discrepant supplier.	In 137 cases (18.0%), ChEMBL was correct while either the target or the species had to be corrected in the remaining 627 cases.

Louisa Bellis and Yvonne Light. ChEMBL team. European Bioinformatics Institute.

Summary

- By comparing bioactivities three database suppliers have extracted from the same source article, we were able to identify discrepancies and calculate error rate estimates.
  - Error rate estimates vary by parameter:
    ligand > target > value > type
- Validation of the approach shows that it identifies an incorrectly curated value ~65-80 % of the time.
- Database suppliers have been notified of discrepancies in their respective databases for re-curation.
  - thousands of data points have already been corrected in the ChEMBL database
  - similar work is being undertaken by companies representing the WOMBAT and Liceptor databases
- Users of bioactivity data are encouraged, if possible, to double-check the data from the original source.

Acknowledgements

Lutz Franke

Louisa Bellis

Yvonne Light