Complementarity Between Public and Commercial Databases of Bioactive Compounds: Extending the Linkage Between Chemistry and Biology

International Conference on Trends for Scientific Information Professionals, Nice, October 2008

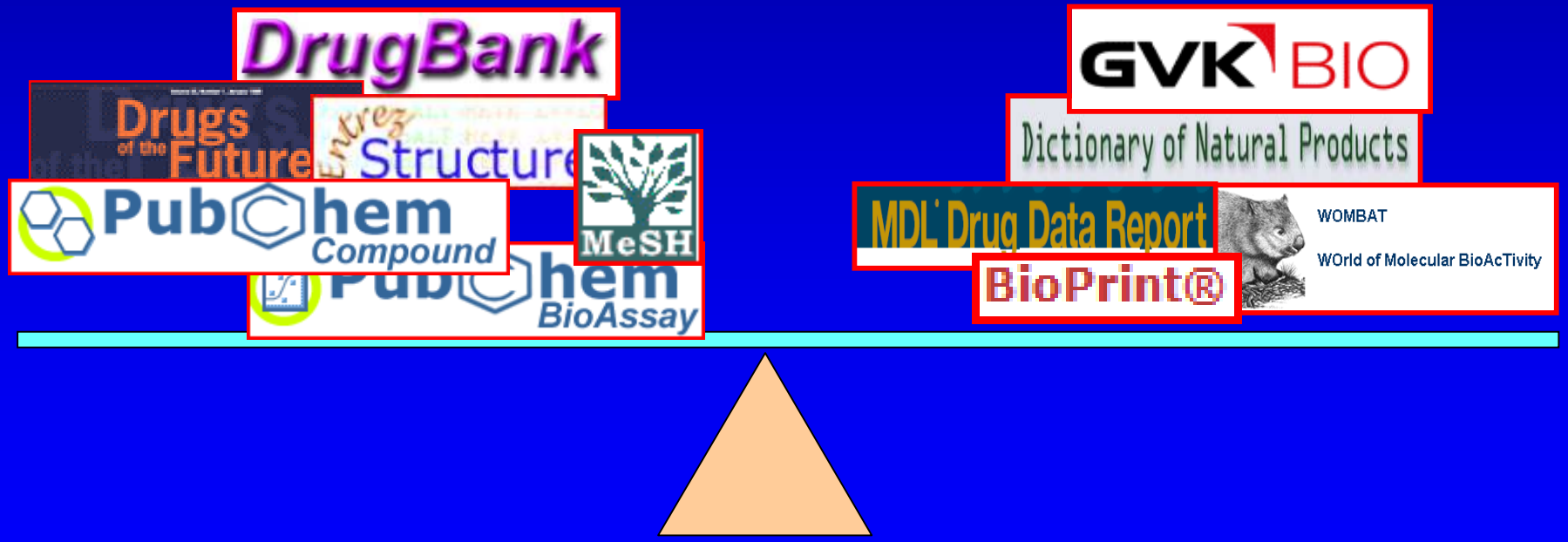
Christopher Southan,

ChrisDS Consulting
Now at EMBL-European Bioinformatics Institute
Welcome Trust Genome Campus, Cambridge, UK)

Assesing the Balance in Content and Bioactivity Linking

Public

Commercial



The Value of Explicit Compound-to-sequence Links

…document "U" includes assay description "V" and a result "W" that defines compound "X" as an activity modulator of protein "Y" with sequence "Z"….

Expert curation of U-V-W-X-Y-Z relationships > databases

WOMBAT

WOrld of Molecular BioAcTivity

~ 130,000 cpds, ~1,300 sequences, ~7,000 papers

GVK BIO

~ 1.5 million cpds ~ 2,000 sequences ~ 20,000 patents and papers

DrugBank

~ 4,000 cpds, 502 sequences

PubChem Compound

83 protein targets with bioassay data, and ~6,000 cpds in PDB structures

Project Objectives

1. Produce standardised comparisons between public and non-proprietary commercial sources
2. Include databases or subsets with explicit chemistry-target or other types of bioactive links
3. Review similarities and differences in content

Methods

1. Normalise downloaded sources by removing fragments
2. Derive canonical tautomer
3. Generate, compare and retain unique molecular hashcodes
1. Prepare “all-against-all” content overlap matrix
2. Perform selected merges and Venn-type comparisons

AstraZeneca

Post-filtration Compound Counts (Oct 2006)

GVKBio	1,488,288
GVKBio Journals	542,858
GVKBio Patents	1,034,548
GVKBio Drug	1,933
WOMBAT	128,120
PubChem	7,268,193
PubChem Prous	3,318
PubChem PDB	5,626
PubChem actives	35,671
PubChem pharmacol	6,070

Bioprint	2,437
ZINC FDA	1,200
DrugBank	3,723
DrugBank small mol	1,018
DrugBank exp drugs	2,737
Dict. Nat.Prod.	132,831
MDDR	159,867
MDDR launched	1,118
CMC	8,189

All-vs-all Result Matrix

	GVKBio	GVKBio Journals	GVKBio Patents	GVKBio DrugDatabase	Wombat	PubChem	PubChem Prous	PubChem PDB	PubChem actives	PubChem pharmacol
GVKBio	1 488 288	542 858	1 034 548	1 663	120 817	439 766	2 563	2 628	9 347	4 221
GVKBio Journals		542 858	89 118	1 637	119 875	245 735	2 424	2 522	9 102	4 090
GVKBio Patents			1 034 548	977	30 135	237 643	1 215	1 052	2 277	2 105
GVKBio DrugDatabase				1 933	683	1 795	733	260	647	1 402
Wombat					128 120	89 305	995	1 114	3 545	1 568
PubChem						7 268 193	3 318	5 626	35 671	6 070
PubChem Prous							3 318	236	541	1 374
PubChem PDB								5 626	538	767
PubChem actives									35 671	1 690
PubChem pharmacol										6 070

GVKBIO

- At just under 1.5 million GVKBIO is divided between journals and patents at approximately 1:2 ratio, with an overlap of 89,000
- GVKBIO covers 93% of WOMBAT and is ~10x larger
- WOMBAT captured over 7000 compounds not found in GVKBIO
- 29% of GVKBIO is represented in PubChem, split evenly between journals and patents
- Includes 25% of cpds reported as active in any of the screening data sets in PubChem and 70% with a pharmacology link in PubChem via MeSH.

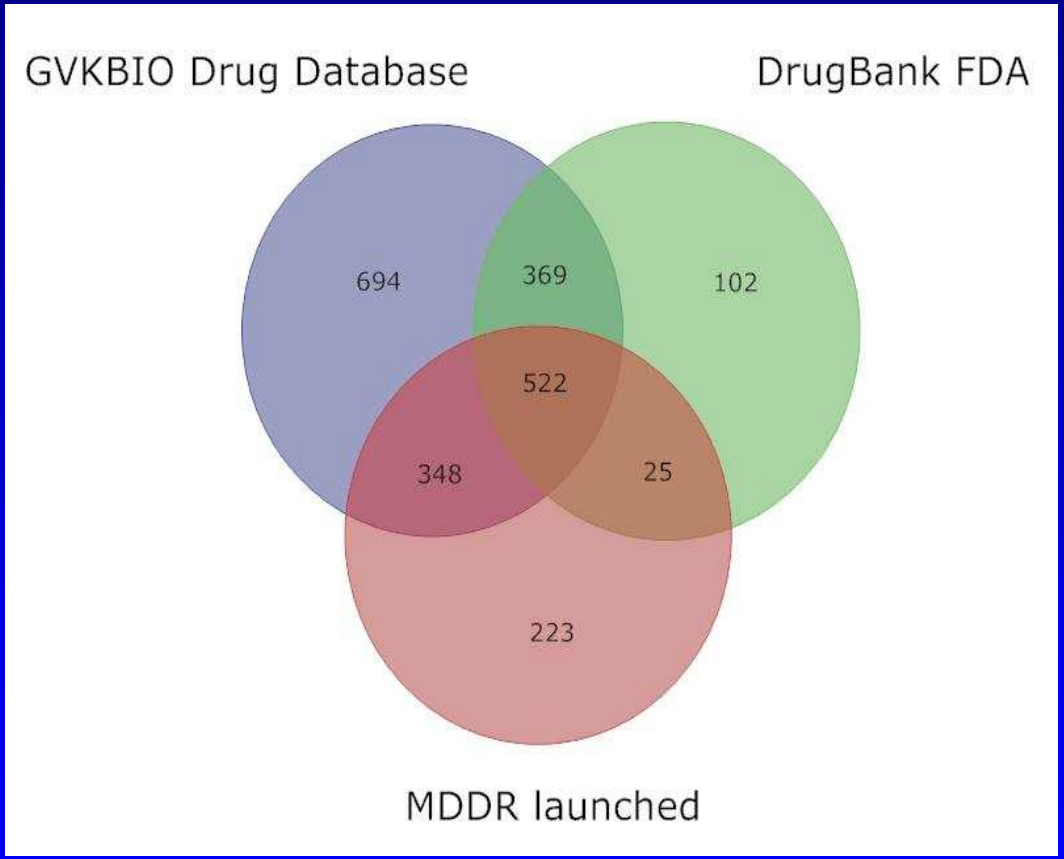
PubChem

- Only 3% screened within the system so far (11% active)
- Largest coverage of every other database, except WOMBAT, of which PubChem covers some 3,000 less compounds than GVKBIO
- 46% of DNP, 42% of MDDR, 92% of DrugBank, 93% of CMC and 95% of BioPrint and MDDR launched
- Covers 0.43 mill of GVKBIO
- GVKBIO patent overlap shows that the number of PubChem compounds with potential claims is 238,000

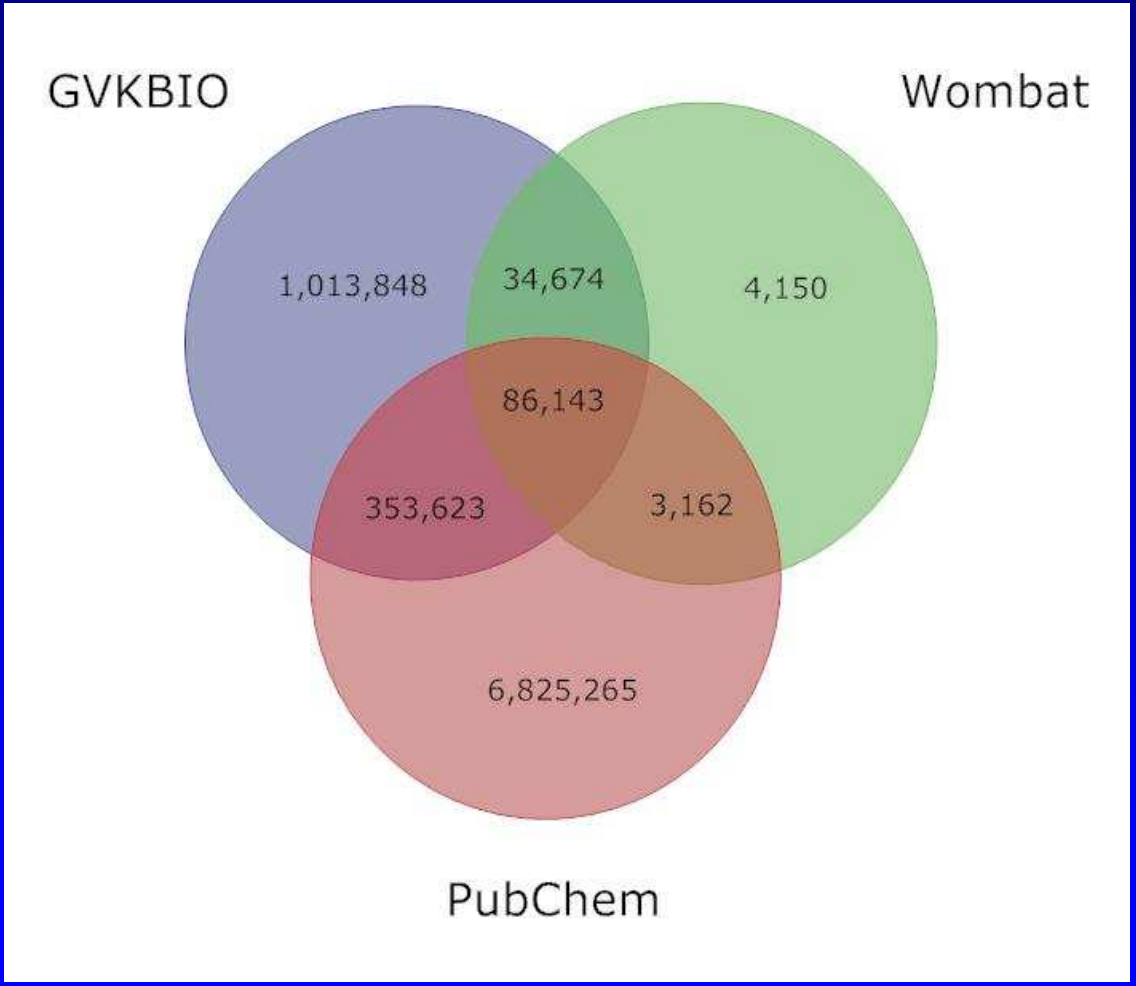
Prous Drugs of the Future

- A review journal for new compounds in development
- 3318 cpds in PubChem with document outlinks (but no inlinks)
- 1374 in PubChem MeSH pharmacology
- Selected overlaps
  - 2,628 in GVKBIO (with document-cpd-sequence links)
  - 733 in GVKBIO Drugs (“ “ “)
  - 994 in WOMBAT (“ “ “)
  - 1,875 in MDDR, 734 in MDDR launched
  - 543 in DrugBank
- Numbers allow inferences on triage through different sources

Comparison of Approved Drug Sets



GVKBIO vs. PubChem vs. WOMBAT

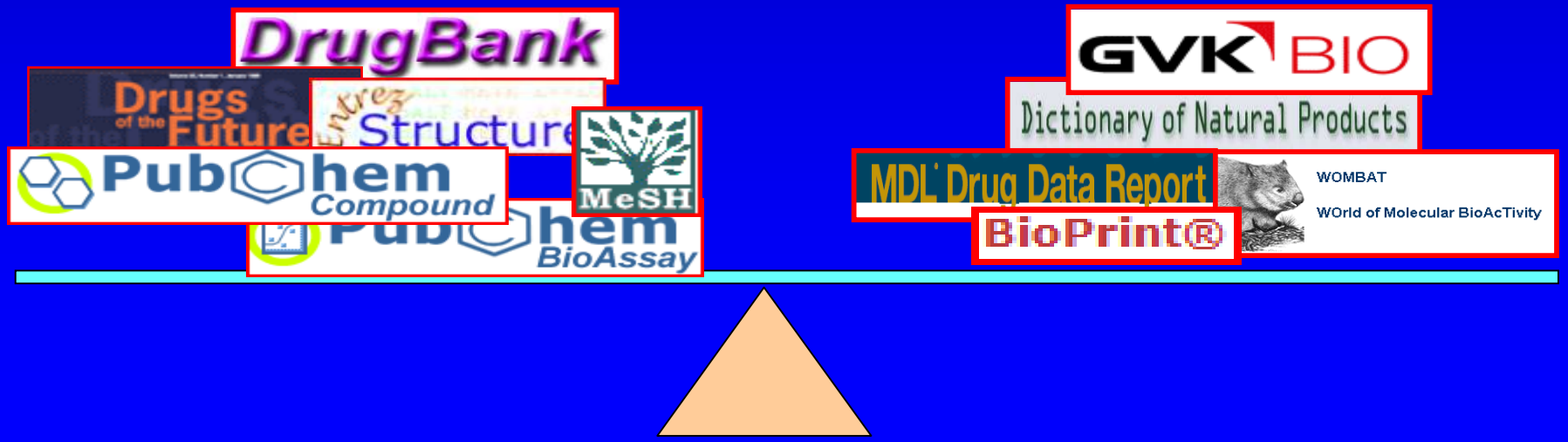


Merge of all Bioactives

- Bioprint, CMC, DNP, DrugBank, GVKBIO, GVKBIO DD, MDDR, PubChem Prous, PubChem PDB, PubChem active, PubChem Pharmacol, ZINK FDA, WOMBAT (not entire PubChem)
- Gives a total merge of **1,976,273**
- Filtered to unique content reduces to **1,741, 392**
- Relatively small redundancy collapse **234, 881** (11%)
- Indicates substantial individual unique content

Summary

- The methods we have developed allow accurate and highly informative comparisons of database content
- Based on coverage *per se* the balance is shifting in the public direction but substantial numbers of bioactives are unique to commercial sources
- Public sources offer different linking and mining functionality that are complementarity to commercial sources
- Compound-assay-result-protein relationships are high value



References and Acknowledgments

- Reference: "Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics" Chris Southan, Péter Várkonyi and Sorel Muresan, *Current Topics In Medicinal Chemistry*, 2007, 7(15), 1502-8

- Thanks to: Tudor Oprea for WOMBAT data