The big data join in pharmacology: linking structures, databases and documents

Christopher Southan, Elena Faccenda, Joanna L Sharman, Simon D Harding, Adam J Pawson, Jamie A Davies

Deanery of Biomedical Sciences, IUPHAR/BPS Guide to PHARMACOLOGY, UK

Introduction: The pharmacological literature and patents connect compound structures to their bioactivity. However, entombing these relationships among millions of PDFs is seriously problematic. The situation is ameliorated by resources that extract data relationships the authors put in to their PDFs back out into structured database records. The IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb) has been doing this by stringent curation of ligands and their quantitative activity against protein targets [1]. Our citations are submitted to PubChem (PC), who then link to PubMed (PM). This study presents an overview of this connectivity.

Methods: For GtoPdb entries in PC Substance we used the PC interface to count our submitted PM links. This gives the PC to PM mapping counts from which we analysed the PM links. We then performed reciprocal analyses (i.e. PM to PC) by selecting PM sets. We then compared two journals by counting structure links by year and source.

Results: From 8988 GtoPdb-submitted substances in PC (release 2017.5), 7309 are linked to 8980 PM entries and 5632 links to chemical structures in PC the rest being antibodies and larger peptides. From the 8980 PMIDs, the Journal of Medicinal Chemistry (JMC) accounted for 1003 as our most frequently cited primary source of structure-to-activity mappings. For the British Journal of Pharmacology (BJP) most of the 345 cross-references were development compounds.

Further analysis showed that from 2014 to 2017 the BJP to PC links of approximately 30 structures per year are mostly from GtoPdb and the Comparative Toxicology Database. However, going back to 2010-12, this increased to 500-800 connections, mainly derived from the IBM automated chemical extraction from abstracts. A similar pattern was observed for JMC.

Conclusion: Navigation between documents and databases is an essential competence for pharmacologists and drug discovery but the NCBI Entrez system is daunting. GtoPdb is a major contributor of high-quality links and provides a first-stop to guide users into the PC and PM systems. However, our results indicated potentially serious specificity issues with automated chemistry-to-journal linking from non-GtoPdb sources.