Multiplexing analysis of 1000 approved drugs across 70 million PubChem entries: Will the correct structures please stand up?

Christopher Southan

IUPHAR/BPS Guide to PHARMACOLOGY, Center for Integrative Physiology, University of Edinburgh

ACS CINF session: The Growing Impact of Big Data in the World of Chemical Information
Abstract

Database molecular entries for approved drugs are the Crown Jewels of over 50 years of global R&D. However, a surprising degree of uncertainty surrounds exact numbers and explicit chemical structures. Choosing a representative approved drug or clinical candidate is becoming harder because of different molecular representations (i.e. structural multiplexing). In this work results will be presented from the analysis a 1000 drug set compiled inside PubChem. This showed that each structure had been submitted (on average) 81 times. In addition the “same connectivity” operator indicted 21 canonically related CIDs and each drug represented in 44 mixtures. We can also detect the “split bioactivity” problem where 135 CIDs related to taxol, 12 have bioassay results. As the totality of public chemical structures pushes towards 100 million we can track a constellation of problems related to the type of statistics above. In particular, the recently increased open availability of patent extracted chemistry and broadening vendor choice is generally welcome by database users. However, analysing entries related to the 1000 drugs across time indicates both types of expansions come with a cost. For example, the 55 million vendor CIDs show increased unresolved chirality (i.e. flat versions) and/or E/Z positions (crossed-bonds). In addition, noticeable “patent-picking”, including mixtures, suggest vendor submissions are increasing in virtual, rather than extant structures. The 21 million automated and manual patent extractions also bring in a variety of artefacts, such as shotgun exemplifications of mixtures, chiral permutations and virtual deuteration. Also, 85% are devoid of BioAssay data links. As a solutions to at least some of these problems, PubChem facilitates particularly effective query selects and filters predicated on their advanced relationship rules. Notwithstanding, the inexorable increase in multiplexing can confound the less experienced and is arguably reaching problematic proportions across all “big data” chemical resources.
Compiling a 1000 drug set

1. Query at the SID level: approved[All Fields] AND "DrugBank"[SourceName] = 1533
2. Find related data < compound < same CID = 1504
3. Select ChEMBL CIDs = 1458721
4. ChEMBL AND DrugBank = 1358
5. .. Covalent unit 1 = 1329
6. AND Therapeutic Target Database (13913) = 1040
7. = 1001 (remove one manually)
8. = 1000
Substances

• 1000 drugs pivoted to 82194 SIDs
• i.e. 82 submissions for every CID
Same connectivity

- = 22633
Outline