editorial

Antony J. Williams

Sean Ekins

A quality alert and call for improved curation of public chemistry databases

In the last ten years, public online databases have rapidly become trusted valuable resources upon which researchers rely for their chemical structures and data for use in cheminformatics, bioinformatics, systems biology, translational medicine and now drug repositioning or repurposing efforts. Their utility depends on the quality of the underlying molecular structures used. Unfortunately, the quality of much of the chemical structure-based data introduced to the public domain is poor. As an example we describe some of the errors found in the recently released NIH Chemical Genomics Center 'NPC browser' database as an example. There is an urgent need for government funded data curation to improve the quality of internet chemistry and to limit the proliferation of errors and wasted efforts.

US funding agencies have been investing in the development of public domain chemistry platforms with the primary attention being given to the informatics platform itself rather the quality of the data content. This is clearly exemplified by the recently released NPC browser from the NIH Chemical Genomics Center (NCGC) [1]. Public online databases such as PubChem, ChemIDPlus [2] and the EPA's ACToR [3], to name just a few, have rapidly become trusted valuable resources which researchers rely on for downloadable chemical structures and associated data. While online chemistry databases can certainly be of value, we feel the reader should be immediately alerted to consider issues of data quality when using these resources and we call into question both their status and the trust we place in them. To our knowledge the issues we raise, using the example of a recently released database, have not been described elsewhere and the user community, and funding agencies, should not ignore them any longer. The development of cheminformatics platforms without due care given to the data quality they contain, is a poor strategy for long term science.

In the last decade numerous attempts have been made to expand our understanding of biological mechanisms by producing vast ligand and protein–protein interaction databases and by the application of computational methods to mine the data and, where possible, develop computational models. These approaches have enabled: the clustering of biological activity spectra similarity

1359-6446/06/$ - see front matter © 2011 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2011.07.007

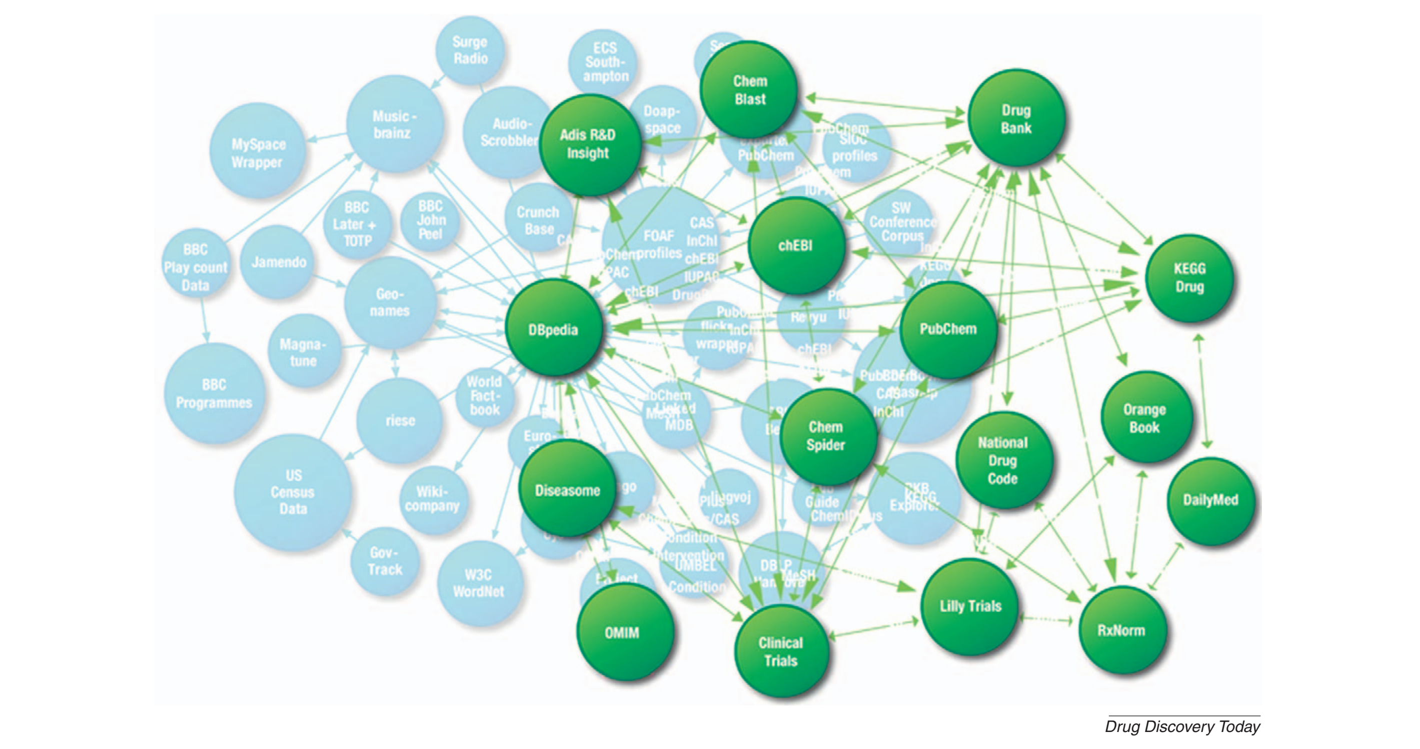
[4–6], global mapping of pharmacological space [7], drug-target networks of approved drugs [8,9], side effect networks from the World Drug Index [10], and prediction of off-target effects of FDA approved drugs and investigational compounds [11], to name but a few of the many examples.

By assimilating various data sources together and meshing data on drugs, proteins and diseases these various databases, network and computational methods may be useful for evaluating molecules for repurposing for new uses that could accelerate drug discovery screening efforts [12–16]. The utility of such databases clearly depends on the quality of the data they contain and, in terms of assembling and interlinking of the databases, on the quality of the underlying molecular structures used.

Virtually all of the databases created are dependent on predominantly freely available sources of molecular structures such as those available from the National Institutes of Health PubChem database [17], the European Bioinformatics Institute-European Molecular Biology Laboratory databases [18], DrugBank [19], the Human Metabolome Database [20], KEGG [21] and ChemSpider [22,23]. Until recently [1], there was no recognized ‘gold standard’ public database of molecules that represents current FDA approved drugs (with validated structures), which failed during clinical trials, were withdrawn by regulatory authorities because of adverse events or were removed from the market by manufacturers.

Observations on compound quality from the NPC browser

We have previously noted that while PubChem has become both a repository of deposited information and a source of trusted information for other databases to use (see Fig. 1), errors have been found to proliferate from here to other databases on the internet when the content is downloaded and reused [24]. We can look at a more recently released public chemistry database as an example of why this is an ongoing issue that needs urgent attention. The NCGC released ‘*a comprehensive resource of clinically approved drugs to enable repurposing and chemical genomics*’ [1] in the form of the NPC browser. This database will be used along with the NCGC screening resources as a component of the NIH therapeutics for rare and neglected diseases (TRND) program. It was suggested that considerable effort went into dealing with semantic errors and sourcing the correct structures [1]. However, from our analysis of the ‘HTS amenable compounds’ subset of data for >7600 compounds it has been possible to identify fundamental errors in stereochemistry (see Fig. 2), valency issues and charge imbalances [25–27]. Similar issues were seen upon analysis of the ‘NPS screening data set’ for >3200 compounds which were suggested as having undergone quality control. Based on early analyses, estimates range from 5% to over 10% of the molecules having errors, while close examination of particular subsets have shown errors of >70% in the absolute structural integrity of the chemicals [28]. These analyses do not consider that the whole dataset consists of over 14 000 compounds and these may also have errors in them requiring correction. While the NCGC suggest the community can help them check the database quality [1], correction of these errors manually will be an enormous undertaking regardless of the much needed curation mechanism, and it is not clear that this will improve the data quality fast enough, or whether there is funding to perform this. In fact, as yet, it is not



Drug Discovery Today

FIGURE 1

A map of linked open data. The green circles and interconnections show various chemistry databases and the links between them showing the sharing of structures between various online resources of value to drug discovery (http://www.w3.org/2009/Talks/1005-jaoo-egp/LODD.png).

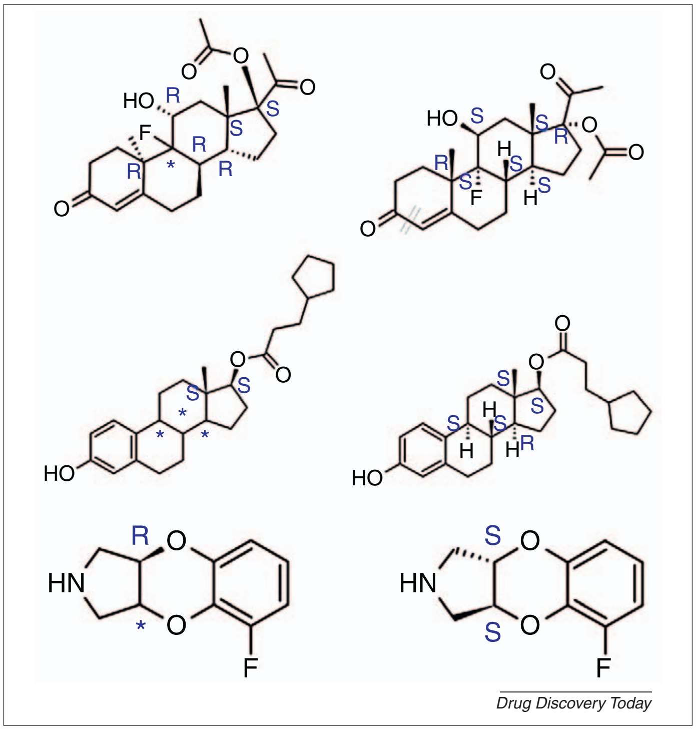


FIGURE 2

Stereochemistry issues with the content of the NPC browser database: left hand side 'NCGC Structures' and right hand side 'Correct Structures'.

widely reported that there are fundamental structure errors in the database. Our observations also raise the issue of whether this database was adequately reviewed before it was published as part of a peer reviewed article, and if not, what reliable mechanisms exist for reviewers to perform this? What if any tools are available to check compound quality of databases? Should molecule databases have some authentication certificate as a sign of quality? It should be noted that the data reported here represent a point in time for the dataset and curation efforts are ongoing.

Chemistry database quality and its effects are themselves little studied. The effects that the correctness of structures has on computational models has been discussed by others [29,30] and indicated that ~10% of datasets used in quantitative structure activity relationship (QSAR) studies published in the Journal of Medicinal Chemistry and in QSAR and Combinatorial Science contain errors with respect to their chemical structures and/or biological activities [29]. While these data may not specifically relate to public chemistry databases but rather peer-reviewed literature, they provide evidence that errors can easily be found and from these publications and that they can proliferate further as that data is cited and reused. The importance of chemical data curation in the context of QSAR modeling is important [31], with small structural errors in a dataset leading to significant losses of predictive abilities of QSAR models. Manual curation of structural data can lead to improvement in the model quality [30,31]. In the field of bioinformatics, the construction of publicly available protein and nucleic acid sequence databases as well as protein structure databases can be impacted by erroneous data which influences the outcome of sequence and structure alignment methods [32].

Recommendations

We propose that it is timely to highlight the issue of quality for internet-based chemistry resources if they and their content are to be used for drug repurposing [33,34] (as in the case of the NPC Browser) or integrated into other cheminformatics or bioinformatics resources for drug discovery. It is not yet appropriate to treat any public chemistry database as authoritative and the errors in the NPC browser suggest that it is not yet '*a comprehensive resource of clinically approved drugs*'. Users should be vigilant in their use and reliance on the chemical structures and data derived from this and other public sources. Ultimately users should obviously be conscious of the trust they are placing in such free resources. There is an urgent need for recognition of this issue and for government funding of data curation for the NPC Browser, PubChem, and other databases to improve the overall quality of chemistry on the internet and stem the proliferation of errors. Without this, crowd-sourced efforts to validate the data in other resources such as ChemSpider and Wikipedia will need to serve the community as fully as possible with limited resources. In the meantime, researchers using previously published data or software resources that leverage any of these databases (that may have erroneous structural content) should either take steps to determine any errors or use the outputs and predictions with extreme caution. Publishers also need to consider the quality of molecule databases prior to publishing papers on them, at least requiring random quality checks, and this may require the development of standards that follow some minimum standard for quality and structural integrity in the same way that we have minimal standards for microarray and other data [35]. Several important questions need addressing such as: how do we ensure that molecule structures are as close to 100% correct as possible?; who corrects the errors in public databases?; how much does it cost to create flawed databases and can the research funding be better spent elsewhere perhaps by outsourcing these efforts? There *is* a need for a free 'gold standard' chemical/drug database, however the NPC browser requires considerable careful curation before reaching that status.

Conflicts of interest

AJW is employed by the Royal Society of Chemistry which owns ChemSpider and associated technologies.

References

1 Huang, R. *et al.* (2011) The NCGC Pharmaceutical Collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci. Transl. Med.* 3, 80ps16

2 Anon., . http://chem.sis.nlm.nih.gov/chemidplus/

3 Judson, R. *et al.* (2009) The toxicity data landscape for environmental chemicals. *Environ. Health Perspect.* 117, 685–695

4 Fliri, A.F. *et al.* (2005) Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* 102, 261–266

5 Fliri, A.F. *et al.* (2005) Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. *J. Med. Chem.* 48, 6918–6925

6 Fliri, A.F. *et al.* (2005) Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* 1, 389–397

7 Paolini, G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* 24, 805–815

8 Wishart, D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36 (database issue), D901–D906

9 Yildirim, M.A. *et al.* (2007) Drug-target network. *Nat. Biotechnol.* 25, 1119–1126

10 Bender, A. *et al.* (2007) Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* 2, 861–873

Editorial

11 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462, 175–181

12 Chong, C.R. and Sullivan, D.J., Jr (2007) New uses for old drugs. *Nature* 448, 645–646

13 Jensen, N.H. and Roth, B.L. (2008) Massively parallel screening of the receptorome. *Comb. Chem. High Throughput Screen.* 11, 420–426

14 Strachan, R.T. *et al.* (2006) Screening the receptorome: an efficient approach for drug discovery and target validation. *Drug Discov. Today* 11, 708–716

15 O'Connor, K.A. and Roth, B.L. (2005) Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat. Rev. Drug Discov.* 4, 1005–1014

16 Roth, B.L. *et al.* (2004) Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.* 102, 99–110

17 Anon., . http://pubchem.ncbi.nlm.nih.gov/

18 Anon., . http://www.ebi.ac.uk/chembldb/index.php

19 Wishart, D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34 (database issue), D668–D672

20 Wishart, D.S. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35 (database issue), D521–D526

21 Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30

22 Anon., . www.chemspider.com

23 Williams, A.J. (2008) Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today* 13, 502–506

24 Williams, A.J. *et al.* (2009) Free online resources enabling crowdsourced drug discovery. *Drug Discov. World* 10, 33–38

25 Williams, A.J. (2011) Reviewing Data Quality in the NCGC Pharmaceutical Collection Browser. ChemConnector Blog, http://www.chemconnector.com/2011/04/28/reviewing-data-quality-in-the-ncgc-pharmaceutical-collection-browser/

26 Williams, A.J. (2011) What is a Drug? Data Quality in the NCGC Pharmaceutical Collection Browser Part 2, http://www.chemconnector.com/2011/05/02/what-is-a-drug-data-quality-in-the-ncgc-pharmaceutical-collection-browser-part-2/

27 Williams, A.J. (2011) Support for Common Compounds in the NPC Browser. Data Quality Part 3, http://www.chemconnector.com/2011/05/02/support-for-common-compounds-in-the-npc-browser-data-quality-part-3/

28 Williams, A.J. (2011) Unreported results. In Preparation

29 Oprea, T. *et al.* (2002) On the propogation of errors in the QSAR literature. *Euro QSAR*

30 Fourches, D. *et al.* (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* 50, 1189–1204

31 Young, D. *et al.* (2008) Are the chemical structures in your QSAR correct? *QSAR Comb. Sci.* 27, 1337–1345

32 Thompson, J.D. *et al.* (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* 6, e18093

33 Ekins, S. *et al.* (2011) In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discov. Today* 16, 298–310

34 Ekins, S. and Williams, A.J. (2011) Finding promiscuous old drugs for new uses. *Pharm Res.* 28 (8), 1785–1791

35 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365–371

Antony J. Williams
Royal Society of Chemistry, 904 Tamaras Circle,
Wake Forest, NC 27587, USA
williams@rsc.org

Sean Ekins
Collaborations in Chemistry, 601 Runnymede Avenue,
Jenkintown, PA 19046, USA
Department of Pharmaceutical Sciences, University of Maryland,
MD 21201, USA
ekinssean@yahoo.com