Curation of chemogenomics data

To the Editor: With the rapid accumulation of data in all areas of chemical biology research, scientists rely increasingly on historical chemogenomics data and computational models to guide small-molecule bioactivity screens and chemical probe development. However, there is a growing public concern about the frequent irreproducibility of experimental data reported in peer-reviewed scientific publications[1,2]. An editorial in this journal[3] emphasized a critical need to address this problem, an issue that has also received attention from the US National Institutes of Health (NIH) leadership[4]. Since successful development of chemical probes and robust screening assays—one central objective of chemical biology—rely on the prior art in the field, it is critical that researchers establish the highest possible quality standards for data deposited in chemogenomics databases.

Concerning the impact of poor data in chemogenomic databases, we[5] and others[6] have shown that inaccurate and inconsistent representations of chemical structures in available molecular datasets result in models of poor accuracy, whereas data curation improves the modeling outcome. Researchers relying on non-curated historical data are taking a risk of corrupting their results owing to the following 'five I's': data may be incomplete, inaccurate, imprecise, incompatible and/or irreproducible. These considerations emphasize the need for thorough curation as the first critical step of any data analysis study to ensure the stability and reliability of the models and to guide experimental follow-up[5].

As one means of addressing the data quality problem, we propose a general chemical and biological data curation workflow (**Fig. 1**) that relies on existing cheminformatics approaches to flag, and in some cases correct, possibly erroneous entries in large chemogenomics datasets. This workflow begins with chemical data curation following a previously established protocol[5] (step 1 in **Fig. 1**), resulting in the identification and correction of structural errors. Duplicate analysis (step 2) assesses data quality and removes duplicate chemical structures and contradictory records. Analysis of intra- and interlab experimental variability (step 3) and exclusion of unreliable data sources (step 4) help increase data quality and aid decision-making about combination of data from different sources. Detection and verification of activity 'cliffs' (step 5) and calculation and tuning of the dataset modelability index[7] (step 6), which estimates the feasibility of obtaining predictive quantitative structure-activity relationship (QSAR) models for a given dataset, serve as additional indicators of data quality. Consensus QSAR modeling (step 7), used for the identification and correction of potentially erroneous values or categories of compound bioactivities (step 8), conclude the workflow.

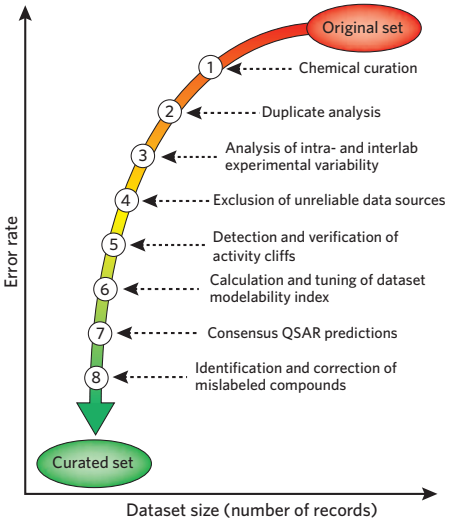

Figure 1 | General workflow for comprehensive curation of chemogenomics datasets. Each step can be done using existing cheminformatics techniques and software tools. The workflow ensures the detection and elimination of the following: nonstandardized and duplicated chemical structures (steps 1 and 2); records associated with unreliable data sources or high experimental variability (steps 3 and 4); structural outliers and unverified activity cliffs (steps 5 and 6). Some mislabeled compounds can thereby be identified and corrected (steps 7 and 8).

As a community, we must take multifaceted approaches to ensure the quality and reproducibility of chemogenomics data through better data generation and reporting. The Nature family of journals[8] have taken steps in this direction by removing space restrictions for method sections and having external statisticians verifying the correctness of statistical tests reported in some manuscripts considered for publication. The NIH is also developing plans to stimulate researchers to enhance reproducibility of their research results (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html). It is also crucial for journals to support and encourage the use of standardized electronic protocols and formats (such as MIABE[9]) for chemical data sharing and to require authors to upload their data electronically to public repositories at the time of manuscript submission.

Among other measures, the chemical biology community should adopt a culture of curation as a mandatory component of primary data processing and a prerequisite for data sharing. Chemical and biological data curation workflows can be developed further and utilized to flag (and where possible, fix) those records and ultimately improve the quality of data analysis and the prediction performances of modeling approaches. Experimental and computational scientists should convene to agree upon standards and best practices for data generation, reporting and curation of chemogenomics data, which will improve data reproducibility and accelerate the progression from data to knowledge in chemical biology research.

Acknowledgments

The authors are grateful for financial support from the US National Institutes of Health (GM 096967 and GM66940) and Environmental Protection Agency (RD 83499901). D.F. thanks the North Carolina State University Chancellor's Faculty Excellence Program.

Competing financial interests

The authors declare no competing financial interests.

References

1. Baker, M. *Nature* **521**, 274–276 (2015).
2. Prinz, F., Schlange, T. & Asadullah, K. *Nat. Rev. Drug Discov.* **10**, 712 (2011).
3. Anonymous. *Nat. Chem. Biol.* **9**, 345 (2013).
4. Collins, F.S. & Tabak, L.A. *Nature* **505**, 612–613 (2014).
5. Fourches, D., Muratov, E. & Tropsha, A. *J. Chem. Inf. Model.* **50**, 1189–1204 (2010).
6. Young, D., Martin, D., Venkatapathy, R. & Harten, P. *QSAR Comb. Sci.* **27**, 1337–1345 (2008).
7. Golbraikh, A., Muratov, E., Fourches, D. & Tropsha, A. *J. Chem. Inf. Model.* **54**, 1–4 (2014).
8. Anonymous. *Nature* **496**, 398 (2013).
9. Orchard, S. *et al. Nat. Rev. Drug Discov.* **10**, 661–669 (2011).

Denis Fourches[1], Eugene Muratov[2] & Alexander Tropsha[2]

[1]Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA.
[2]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.
e-mail: alex_tropsha@unc.edu or dfourch@ncsu.edu

© 2015 Nature America, Inc. All rights reserved.