Genome Medicine

MUSINGS

So what have data standards ever done for us? The view from metabolomics

Julian L Griffin*[1,2,3] and Christoph Steinbeck[4]

Abstract

The standardization of reporting of data promises to revolutionize biology by allowing community access to data generated in laboratories across the globe. This approach has already influenced genomics and transcriptomics. Projects that have previously been viewed as being too big to implement can now be distributed across multiple sites. There are now public databases for gene sequences, transcriptomic profiling and proteomic experiments. However, progress in the metabolomic community has seemed to falter recently, and whereas there are ontologies to describe the metadata for metabolomics there are still no central repositories for the datasets themselves. Here, we examine some of the challenges and potential benefits of further efforts towards data standardization in metabolomics and metabonomics.

Whatever the branch of genome science one is part of, the need for data standards (more specifically, standardized ways to describe an experiment) and central repositories for the huge multivariate datasets that researchers are now acquiring seem self-evident. The community needs to be able to reproduce analyses for key experiments, and if the experimenter is satisfied with the quality of the data, why should the data not be made available for all?

In many ways, central repositories for data made the field of genome science accessible to the wider academic community, with over 192 complete sequenced genomes now available for researchers to interrogate. Similarly, moving from genome sequencing to functional genomics, the Microarray Gene Expression Databases (MGED) society developed a community-wide agreement on reporting microarray data (the Minimum Information About a Microarray Experiment or MIAME) and, in conjunction with the European Bioinformatics Institute (EBI) and the National Institutes of Health (NIH), the community developed databases to house the experimental data generated by microarray experiments, such as The ArrayExpress Archive and the Gene Expression Omnibus [1,2]. It was no accident that these developments occurred in parallel, because there is first a need to define a standard reporting language (ontology) before the creation of a database. As well as the 'carrot' of a community-wide resource, the MGED society was also incredibly successful at getting journals on board to police the deposition of data.

The next logical extension for functional genomics after MIAME was to extend these developments from the transcriptomics community to proteomics. The Human Proteome Organization (HUPO), an international consortium of industry, academic and government scientists, set out to extend standard reporting to proteomics. Just as in the field of microarray experiments, developments in data standardization also led to the construction of repositories. The Proteomics Identifications (PRIDE) database is a centralized public repository for proteomic data. The aim of the database is to provide the proteomics community with the ability to store data on protein and peptide expression and also the associated data describing the identifications. More recently, it has expanded to also capture information on post-translational modifications. PRIDE was developed through a collaboration between the EBI and Ghent University, Belgium, and its development has since been closely linked with the HUPO Proteomics Standardization Initiative.

So with transcriptomics and proteomics being such success stories for data standardization and deposition, it seemed logical to extend this to metabolomics. This seemed to be a relatively straightforward process, given that there were already several examples of 'metabolic databases' that contained metabolomics data in all but name. Before the coining of the words metabolomics and metabonomics there were already databases for recording chemical shift and coupling patterns of small molecules, largely to assist chemists and biochemists in mixture

*Correspondence: jlg40@cam.ac.uk
[1]The Department of Biochemistry, Tennis Court Road, University of Cambridge, Cambridge CB2 1GA, UK
Full list of author information is available at the end of the article

analysis (for example, NMRShiftDB for organic structures [3], and BioMagResBank [4], initially for nuclear magnetic resonance (NMR)-determined protein structures, but then extended to small organic molecules and now encompassing metabolomic data too). Data exchange formats for NMR datasets are available from both the Collaborative Computing Project for NMR (CCPN) project, which offers a data model for macromolecular NMR and related areas, and the Joint Committee on Atomic and Molecular Physical Data (JCAMP-DX) [5,6]. Similar developments have also occurred in mass spectrometry, and mass-spectrometry-based metabolomics also benefits from some similarities with proteomic analyses.

Thus, following various publications on how metabolomic experiments should be described - such as the Minimum Information about a Metabolomics Experiment (MIAMET) and Architecture for Metabolomics (ArMet) [7], which were both written from a plant metabolomics perspective, and the Standardization of Reporting Methods for Metabolic Analysis (SMRS) [8], focusing on NMR-based methods and toxicology and animal functional genomics experiments - it seemed that the time was right to develop a community-wide agreed description of reporting a metabolomics experiment. In 2005 two meetings were held, one in Europe through the EBI and the Metabolic Profiling Forum and one in the USA through the NIH, which served as inputs to the Metabolomics Standards Initiative (MSI) that is orchestrated by the Metabolomics Society [9]. This culminated with the publication of several descriptions in *Metabolomics*, the Society's journal, and one in *Nature Biotechnology* [10] in 2007.

However, here is where the good news begins to falter. Despite it being nearly 3 years since the descriptions were published, there is still a very small number of actual studies that make their data available, and even fewer in a format that would comply with the MSI descriptions [11,12]. Indeed, a quick glance across the MSI descriptions shows that there is no unifying description, and instead a user must define first which description is most appropriate to them depending on what biological system they work on. So why is the metabolomics community different from other communities?

The first answer might be that it is intrinsically more difficult to describe a metabolomic experiment than a transcriptomic or proteomic experiment. The field of metabolomics is dominated by two very different technologies, NMR spectroscopy and mass spectrometry, as well as a variety of other approaches, so producing a standardized workflow is difficult. This is further complicated by the fact that many in the community do not report true concentrations but rather relative intensities; in many cases these equate to a relative concentration, but this does raise the question of how one compares results from an NMR spectrometer with those produced by a mass spectrometer.

There has also been the objection that metabolomics experiments are innately too difficult to explain. There have been many reports of relatively minor changes to components of an experiment producing a big change on the metabolome of an organism. In metabolomic studies in mammalian physiology this has included the effects of altered batches of standard chow, the impact of gut microflora changes from different animal facilities (even within the same facility but in different rooms) and even the impact of loud music on the urinary profiles of mice and rats! In an ideal world a database must capture all this information but clearly this is not feasible. However, these problems face any data standard and this will not just affect metabolomics, but also be a problem for databases from other -omic technologies.

However, there are some positive news stories from metabolomics. Firstly, although there is still a lack of community repositories for data themselves, there are databases for standard NMR and mass spectra, including the Human Metabolome Database [13] and the Madison Metabolomics Consortium Database [14]. There are also databases that are already in use, albeit not across the whole community. The INTERPRET database [15] has been used for several years to distinguish different brain tumors in magnetic resonance spectra collected *in vivo* and the COMET database [16] has demonstrated how metabonomics can be applied to the drug safety assessment field. Finally, there are some metabolomes that urgently need their own database. Although much effort has been expended on developing a description of the human metabolome [17], it is much easier to generate a complete metabolomic description for some other organisms. The yeast metabolome has provided an important research tool for understanding how the network of metabolism is regulated, and a large number of yeast mutants have also been metabolically profiled. Likewise, no obese *Caenorhabditis elegans* model seems to be publishable without a profile of the total fatty acids present, and thus it seems a database of *C. elegans* metabolic changes associated with mutations would be a worthy community resource.

So what can be done? As a community we need to start making our data available, not just for the purposes of the review process but in order to make the raw material accessible for the next generation of metabolomic software and bioinformatics analysis tools, which again can only be developed and optimized if there are data to work with. We also need to start to build descriptions up for key organisms. When manuscripts are reviewed, we as reviewers and editors have to start to ask to see the data, if only to guarantee their quality. Here, journals

themselves can help by providing both a carrot in the form of suitable facilities for supplementary data and a stick in the form of a journal requirement for the raw data. However, the ultimate responsibility must lie with the community. Perhaps the question is not what standards can do for you, but what you can do for data standards!

Abbreviations

EBI, European Bioinformatics Institute; HUPO, Human Proteome Organization; MGED, Microarray Gene Expression Database; MIAME, Minimum Information About a Microarray Experiment; MSI, Metabolomics Standards Initiative; NIH, National Institutes of Health; NMR, nuclear magnetic resonance; PRIDE, proteomics identifications.

Competing interests

JG and CS are recipients of a BBSRC grant entitled MetaboLights to develop a central repository and curated resource for metabolomic data.

Author details

[1]The Department of Biochemistry, Tennis Court Road, University of Cambridge, Cambridge CB2 1GA, UK. [2]The Cambridge Systems Biology Centre, Tennis Court Road, University of Cambridge, Cambridge CB2 1GA, UK. [3]The MRC Centre for Obesity and Related Diseases (MRC CORD), the University of Cambridge Metabolic Research Laboratories, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [4]The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK.

Published: 24 June 2010

References

1. The ArrayExpress Archive [http://www.ebi.ac.uk/microarray-as/ae/]
2. Gene Expression Omnibus [http://www.ncbi.nlm.nih.gov/geo/]
3. NMRShiftDB [http://www.nmrshiftdb.org/]
4. Biological Magnetic Resonance Data Bank [http://www.bmrb.wisc.edu/]
5. The Collaborative Computing Project for NMR [http://www.ccpn.ac.uk/ccpn]
6. IUPAC Committee on Printed and Electronic Publications; Subcommittee on Electronic Data Standards [http://www.jcamp-dx.org/]
7. Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, Taylor J, Fiehn O, Goodacre R, Bino RJ, Hall R, Kopka J, Lane GA, Lange BM, Liu JR, Mendes P, Nikolau BJ, Oliver SG, Paton NW, Rhee S, Roessner-Tunali U, Saito K, Smedsgaard J, Sumner LW, Wang T, Walsh S, Wurtele ES, Kell DB: **A proposed framework for the description of plant metabolomics experiments and their results.** *Nat Biotechnol* 2004, **22:**1601-1606.
8. Lindon JC, Nicholson JK, Holmes E, Keun HC, Craig A, Pearce JT, Bruce SJ, Hardy N, Sansone SA, Antti H, Jonsson P, Daykin C, Navarange M, Beger RD, Verheij ER, Amberg A, Baunsgaard D, Cantor GH, Lehman-McKeeman L, Earll M, Wold S, Johansson E, Haselden JN, Kramer K, Thomas C, Lindberg J, Schuppe-Koistinen I, Wilson ID, Reily MD, Standard Metabolic Reporting Structures working group, *et al.*: **Summary recommendations for standardization and reporting of metabolic analyses.** *Nat Biotechnol* 2005, **23:**833-838.
9. The Metabolomics Standards Initiative [http://msi-workgroups.sourceforge.net/]
10. The MSI Board Members: Sansone SA, Fan T, Goodacre R, Griffin JL, Hardy NW, Kaddurah-Daouk R, Kristal BS, Lindon J, Mendes P, Morrison N, Nikolau B, Robertson D, Sumner LW, Taylor C, van der Werf M, van Ommen B, Fiehn O: **The metabolomics standards initiative.** *Nat Biotechnol* 2007, **25:**846-848.
11. Fiehn O, Wohlgemuth G, Scholz M, Kind T, Lee do Y, Lu Y, Moon S, Nikolau B: **Quality control for plant metabolomics: reporting MSI-compliant studies.** *Plant J* 2008, **53:**691-704.
12. Atherton HJ, Gulston MK, Bailey NJ, Cheng KK, Zhang W, Clarke K, Griffin JL: **Metabolomics of the interaction between PPAR-alpha and age in the PPAR-alpha-null mouse.** *Mol Syst Biol* 2009, **5:**259.
13. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia J, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P, Fang L, Peng J, Fradette R, Cheng D, Tzur D, Clements M, Lewis A, De Souza A, Zuniga A, Dawe M, *et al.*: **HMDB: a knowledgebase for the human metabolome.** *Nucleic Acids Res* 2009, **37:**D603-D610.
14. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, Markley JL: **Metabolite identification via the Madison Metabolomics Consortium Database**. *Nat Biotechnol* 2008, **26:**162-164.
15. Tate AR, Majós C, Moreno A, Howe FA, Griffiths JR, Arús C: **Automated classification of short echo time in in vivo 1H brain tumor spectra: a multicenter study.** *Magn Reson Med* 2003, **49:**29-36.
16. Ebbels TM, Keun HC, Beckonert OP, Bollard ME, Lindon JC, Holmes E, Nicholson JK: **Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: the consortium on metabonomic toxicology screening approach.** *J Proteome Res* 2007, **6:**4407-4422.
17. HUSERMET: Human Serum Metabolome in Health and Disease [http://www.husermet.org/]

doi:10.1186/gm159
Cite this article as: Griffin JL, Steinbeck C: So what have data standards ever done for us? The view from metabolomics. *Genome Medicine* 2010, **2:**38.