Metabolomics (2014) 10:350–353
DOI 10.1007/s11306-014-0656-8

OPINION

Metabolite identification: are you sure? And how do your peers gauge your confidence?

Darren J. Creek · Warwick B. Dunn · Oliver Fiehn · Julian L. Griffin · Robert D. Hall · Zhentian Lei · Robert Mistrik · Steffen Neumann · Emma L. Schymanski · Lloyd W. Sumner · Robert Trengove · Jean-Luc Wolfender

1 The identification challenge

Metabolomics is still faced with several significant challenges which currently limit its full scientific potential. The identification of metabolites is essential to convert analytical data into meaningful biological knowledge. However, identification confidence can vary widely because the process of identification is complex and dependent on the analytical platform and robustness of the methods applied, as well as the databases and resources used. Confident and unequivocal structure identification requires significant effort, which is multiplied dramatically in non-targeted metabolomics studies where 10–100s of metabolites can be deemed as biologically important and require identification. Mass spectrometry (MS), nuclear magnetic resonance spectroscopy (NMR) or integrated MS–NMR strategies (Dunn et al. 2013; Kind and Fiehn 2010; van der Hooft et al. 2011) provide much information for the identification of metabolites (e.g. 1D/2D-NMR and MS/MS).

Relatively high confidence can only be obtained when comparing multiple physicochemical properties of an authentic pure chemical standard to those of the metabolite of interest observed under identical analytical conditions. However, isomers are a significant problem as they often behave differently in biological systems but similarly in analytical platforms with respect to mass, chromatographic retention time and NMR chemical shift. Chiral chromatography or judicious alterations of the column chemistry can help, but is typically not performed for general profiling. Even high resolution-accurate mass (HRAM), MS/MS and retention time comparisons with an authentic standard do not always lead to an unequivocal identification, especially as available standards and databases do not cover the full extent of biochemical diversity. It is often

This work is supported by the Metabolite Identification Task Group of the Metabolomics Society. All authors have contributed equally to this work.

D. J. Creek
Monash University, Parkville, VIC, Australia

W. B. Dunn (✉)
University of Birmingham, Edgbaston, Birmingham, UK
e-mail: w.dunn@bham.ac.uk

O. Fiehn
University of California Davis, Davis, CA, USA

J. L. Griffin
University of Cambridge, Cambridge, UK

R. D. Hall
Plant Research International and Wageningen University, Wageningen, The Netherlands

Z. Lei · L. W. Sumner
The Samuel Roberts Noble Foundation, Ardmore, OK, USA

R. Mistrik
HighChem, Ltd., Bratislava, Slovakia

S. Neumann
Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany

E. L. Schymanski
Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

R. Trengove
Murdoch University, Murdoch, WA, Australia

J.-L. Wolfender
University of Geneva, Geneva, Switzerland

J.-L. Wolfender
University of Lausanne, Lausanne, Switzerland

impractical to rule out all structurally similar candidates experimentally. Literature, database, genome and chemotaxonomic knowledge can be used to propose the most 'biologically likely' isomer, but this approach may be misleading.

So how confident are you in the metabolite(s) you are reporting in the scientific literature and at conferences? Is the metabolite leucine or isoleucine? Are *trans* fatty acids actually *cis* fatty acids? Is PC(18:2/18:2) perhaps PC(18:0/18:4)? The uniqueness, or not, of the identification is often overlooked as the presence of isomers is common. Communicating the confidence or uncertainty in proposed metabolite identification is essential to avoid misinterpretation.

2 Current reporting standards for metabolite identification

In 2007, the Chemical Analysis Working Group (CAWG) of the Metabolomics Standards Initiative (MSI) recommended minimum reporting standards for chemical analysis in metabolomics (Sumner et al. 2007). This was a community effort to build a general consensus on reporting standards in metabolomics. These reporting standards recommended that 'Authors should clearly differentiate and report the level of identification rigor for all metabolites reported' based on a four-level system ranging from Level 1 (identified compound) via Levels 2 and 3 (putatively annotated compounds and compound classes) to Level 4 (unidentified or unclassified metabolites which nevertheless can be differentiated based upon spectral data). Other scientific disciplines apply other rules for reporting chemical identification. For example, in the natural products research community a compound submitted for publication may require data collected from HRAM MS, 1D-NMR (^{1}H and ^{13}C) and additional circular dichroism and/or alpha D values to define the stereochemistry clearly. However, these requirements are generally applied to novel natural products identified for the first time in the literature.

Reporting standards for metabolite identification are essential to integrating metabolomics data within other omics disciplines. This is particularly important in this era of computational biology and large open-access databases, where many end-users will not be (bio)chemists. Although 7 years have passed since the first reporting standards were made available, the application of these simple reporting standards in scientific publications is still limited (Salek et al. 2013). It is essential to use accurate metabolite reporting standards to allow the wider scientific community to assess and interpret data with confidence. Authors must describe the process and level/confidence of metabolite identification in all scientific publications and these should be verified in the review process. It is clear that guidance and education on the benefits and use of reporting standards is required.

3 The Metabolite Identification task group of the Metabolomics Society

The Metabolomics Society aims to support the growth of metabolomics internationally. The society has commissioned a Metabolite Identification task group to work with the community including researchers, community resources such as metabolomics data repositories and instrument suppliers, and build consensus on identification reporting standards.

This international task group has five objectives:

(a) Assess, build consensus and further develop reporting standards with input from the metabolomics community and the data standards task group of the Metabolomics Society.
(b) Educate the metabolomics community on appropriate reporting standards for metabolite annotation and identification.
(c) Ensure application of these standards in international efforts including publishing and data repositories.
(d) Provide information on current tools, resources and workflows, and a forum to discuss their application and applicability.
(e) Provide the opportunity for (i) inter-laboratory comparisons of methods [e.g. CASMI (Schymanski and Neumann 2013)] and (ii) validation of methods applied for metabolite identification in association with other societies and organisations.

The necessity for metabolite identification reporting standards is clear, and we reach out to the community for help to modify the current reporting standards to ensure optimal relevance for the community.

4 Upgrading reporting standards—a community consensus

Previously in (Sumner et al. 2007) it was stated 'The exact basis for what constitutes a valid metabolite identification is still currently debated in the community and a consensus is still evolving'. These reporting standards were an important first stage in an ongoing process. It is now time to reassess the current reporting standards for metabolite identification and identify any changes required because of recent developments in standards initiatives internationally (for example, the Metabolomics Society Data Standards task group and EU-project COSMOS; http://www.cosmos-

fp7.eu). Further, the original recommendations dividing confidence into four levels were highly appropriate and simple to use but focused mainly on MS and NMR studies. As our understanding of the different identification processes increases, the need to clarify these reporting standards has arisen.

Many users of the current four levels apply these because of their simplicity and state this as a significant advantage. However, in its simplicity, it lacks the ability to separate many special cases with similar but not unique levels of identification (e.g. isomers). In most cases there is a wide range of confidence levels associated with each MSI level and we propose to provide sub-levels within the 4 MSI levels as one option which will provide greater distinctions within the existing level system. Consideration will be needed when including the numerous MS and NMR methods available which do not have the same differentiation power. For example, an accurate mass measurement has greater differentiation power than a nominal mass measurement and a 2D NMR spectrum has greater differentiation power than a 1D NMR. As one example, a 5-tiered level system was recently proposed for high mass accuracy and resolution MS/MS data (Schymanski et al. 2014) to move towards a more data-dependent system that distinguishes between a confirmed structure, probable structure, tentative and structurally similar candidates, unequivocal molecular formula and exact mass of interest. A revision of the current 4 MSI levels could make these similarly applicable.

A second option is a quantitative scoring system as an alternative to a revised or upgraded confidence level system as discussed above. For example, The EU Guideline 2002/657/EC (http://old.eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:221:0008:0036:EN:PDF) provides an identification point (IP) system for reporting target compounds (e.g. HR-MS1 + RT = 2 IPs, HR-MS1 + 2 HR-MS/MS fragments + RT = 4.5 IPs where HR = high mass resolution and RT = retention time). Such a quantitative scoring system would need to be expanded to include all technologies applied in metabolomics. The system could also be subjective and become very complex and prone to manipulation when you consider the large number of analytical methods available and their variation in differentiation power. A quantitative IPs system should also include some valuation based upon spectral match scoring where possible. Such scoring systems have long been used in mass spectral matching of experimental data to that of authentic standard spectra in libraries [for example, AMDIS for small molecules (Stein 1999) and MASCOT for proteins (Perkins et al., 1999)].

A third option is to use a quantitative scoring system to enhance the current or revised MSI levels, where the levels could indicate the confidence in the structure and the scores represent a summary of the evidence behind it. Thus, confirmed structures that are present in low concentrations with matrix interferences could be Level 1 but low score (e.g. IP = 2) while isomers that cannot be separated but with good experimental evidence could be Level 3 but IP = 4.5. A rewording of the MSI levels would be necessary to achieve this, but this would introduce a flexible system where users could report score, level or both and, furthermore, allow compatibility with automated methods of identification and reporting.

5 A call to the metabolomics community

It is essential that the metabolomics community as a whole understands and actively adopts the current and future versions of metabolomics reporting standards. Through this article we propose three options and ask the community for feedback. Future developments, though not discussed here in detail, will focus on defining expectations and content for metabolomics supporting information, reporting ambiguous identifications, formats for data integration and data repositories. The Metabolomics Society is hosting a metabolite identification task group forum dedicated to an open discussion available at http://interest-groups.metabolomicssociety.org/. We invite all to provide their comments on metabolite identification, and importantly, your views on the proposed revisions described above.

References

Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., et al. (2013). Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics, 9*, S44–S66.

Kind, T., & Fiehn, O. (2010). Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews, 2*, 23–60.

Perkins, D. N., Pappin, D. J. C., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis, 20*, 3551–3567.

Salek, R. M., Steinbeck, C., Viant, M. R., Goodacre, R., & Dunn, W. B. (2013). The role of reporting standards for metabolite annotation and identification in metabolomic studies. *Gigascience, 2*, 13.

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., et al. (2014). Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environmental Science and Technology, 48*, 2097–2098.

Schymanski, E. L., & Neumann, S. (2013). CASMI: and the winner is …. *Metabolites, 3*, 412–439.

Stein, S. E. (1999). An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry, 10*, 770–781.

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., et al. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics, 3*, 211–221.

van der Hooft, J. J., Mihaleva, V., de Vos, R. C., Bino, R. J., & Vervoort, J. (2011). A strategy for fast structural elucidation of metabolites in small volume plant extracts using automated MS-guided LC-MS-SPE-NMR. *Magnetic Resonance in Chemistry, 49*(Suppl 1), S55–S60.