Rethinking Mass Spectrometry-Based Small Molecule Identification Strategies in Metabolomics

Fumio Matsuda1,2

1Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, 1–5 Yamadaoka, Suita, Osaka 565–0871, Japan
2RIKEN Center for Sustainable Resource Science, 1–7–22 Suehiro-cho, Tsurumi-ku, Yokohama 230–0045, Japan

The CASMI 2013 (Critical Assessment of Small Molecule Identification 2013, http://casmi-contest.org/) contest was held to systematically evaluate strategies used for mass spectrometry-based identification of small molecules. The results of the contest highlight that, because of the extensive efforts made towards the construction of databases and search tools, database-assisted small molecule identification can now automatically annotate some metabolite signals found in the metabolome data. In this commentary, the current state of metabolite annotation is compared with that of transcriptomics and proteomics. The comparison suggested that certain limitations in the metabolite annotation process need to be addressed, such as (i) the completeness of the database, (ii) the conversion between raw data and structure, (iii) the one-to-one correspondence between measured data and correct search results, and (iv) the false discovery rate in database search results.

Please cite this article as: F. Matsuda, Rethinking Mass Spectrometry-Based Small Molecule Identification Strategies in Metabolomics, Mass Spectrom (Tokyo) 2014; 3(3): S0038; DOI: 10.5702/massspectrometry.S0038

Keywords: small molecule identification, metabolomics, database, false discovery rate

INTRODUCTION

I contributed to CASMI 2013 by preparing the challenges and scoring the answers. The participants achieved good scores showing that, because of the extensive efforts towards the development of informatics tools, some of the metabolite signals in the metabolome data can now be automatically annotated by database-assisted small molecule identification. However, the performance of the automated methods was not equal to that of manual determination by the winning team of A. Newsome and D. Nikolic, suggesting that the database-assisted annotation strategy needs to be reconsidered.

In mass spectrometry-based metabolomics, three physicochemical properties or data produced from mass spectrometry are available for structural elucidation: retention times on a chromatogram, mass-to-charge (m/z) ratio of the ion, and fragmentation data from mass or tandem mass (MS/MS) spectra.1,2 Metabolite annotation involves searching databases to find metabolite entries whose physicochemical properties are identical or similar to that of the query.3–5 To highlight the characteristics of metabolite annotation, the search procedure was compared to those used in transcriptomics and proteomics (Fig. 1). In transcriptome analysis by RNaseq, many short reads of cDNAs are obtained using next generation sequencing (NGS), where raw data such as color images are automatically converted to corresponding sequences.6,7 Each read is mapped onto the genome or to the coding sequence. The whole genome sequence is a ‘complete’ experimentally determined database that always has an exact search result (mapping position) for any short sequence that is derived from a target organism. The number of possible short reads that are produced by NGS from a target organism is, in theory, countable or finite. Furthermore, there is a one-to-one correspondence because many short reads can be mapped onto a single position in the genome. When genome information is unavailable, sequences of full-length cDNA that are produced by the de novo assembly of NGS data are used. The functions of these sequences are estimated from a homology search of a database such as GenBank.8 The homology between two sequences is evaluated in terms of the p-value by a probability-based method [such as in the basic local alignment search tool (BLAST)].9,10 The false discovery rate (FDR) in gene annotation results can be controlled by selecting a suitable threshold for the search, although it is difficult to reduce FDR according only e-values.

A typical peptide identification procedure in the discovery proteomics is as following (Fig. 1). A trypsin-digested protein sample is analyzed using liquid chromatography–LC-MS/MS to acquire the MS/MS spectra of the peptides. The raw MS/MS spectra are used to search the artificial MS/MS spectra database of peptides and identify the peptides.
A peptide MS/MS spectral database is constructed from the genome information of a target organism by generating artificial MS/MS spectra for all possible peptides using the rules of trypsin digestion and peptide fragmentation. In the case of peptide identification, the artificial spectral database is complete and the number of possible MS/MS spectra is finite, although the number may be very large due to protein modification. The list of identified proteins inevitably includes false-positive hits that are derived from errors in MS/MS analysis. A target-decoy strategy has been established to elucidate the FDR level. This strategy involves the use of a decoy database that comprises the artificial MS/MS spectra of reversed peptides. A rational search threshold may be determined by evaluating the FDR. Because distinct MS/MS spectra are produced from different peptides, a high one-to-one correspondence exists between the MS/MS spectra and the peptide sequences.

The technical summary showed that the identification procedures in transcriptomics and proteomics are based on (i) the completeness of the database, (ii) the conversion between raw data and sequence, (iii) the one-to-one correspondence between data and correct search results, and (iv) the FDR estimation in the database search results. In the following sections, the current state of metabolite annotation is discussed with respect to these points.

**DATABASE COMPLETENESS**

Almost all CASMI participants used giant compound databases to search for compounds such as PubChem (http://pubchem.ncbi.nlm.nih.gov/) and ChemSpider (http://www.chemspider.com/) whose contain experimentally determined structures of natural and synthetic compounds. Although large databases increase the coverage of compounds, a drawback associated with such databases is not well examined. As discussed in the previous section, nucleotide sequences or MS/MS spectra acquired from a target organism form a finite set of measurable sequence and spectra,
because all of them are derived from the genome sequence (Fig. 1). Based on the nature, a complete database is constructed by including all measurable and no unmeasurable nucleotide sequence and predicted MS/MS spectra. For instance, a complete database for the proteome analysis of budding yeast must contain a predicted MS/MS spectrum of the peptide, NVNDVIAPAFVK, derived from Eno1p (protein translated from ENO1 or YGR254W of Saccharomyces cerevisiae). However, this spectrum must be discarded from the database for a proteome analysis of humans, because the peptide, NVNDVIAPAFVK, is only found in budding yeasts, and not in humans (Fig. 2a).

In the case of metabolite annotation, compound databases for searching are incomplete because a complete list of all metabolites that are produced by target organisms is not available. This incompleteness makes it difficult to identify metabolites that have been discarded from the database. The number of metabolites that may be found in a target organism is practically countably infinite, and the identification of all the metabolites inevitably requires a large metabolite database (Fig. 2b). This incompleteness is not a serious drawback for databases of actual MS/MS spectra, simply because the experimentally measured MS/MS spectra simply become insufficient to fully cover metabolite diversity (Fig. 2b). In addition to continuous data acquisition (while maintaining the spectral quality) and sharing efforts as offered by MassBank and the Human Metabolome Database (HMDB), digitalization of the literature available on experimentally measured MS/MS spectral data may help overcome this bottleneck.

On the other hand, a molecular formula search using the exact mass data circumvents the issue of incompleteness. Biological metabolite databases such as HMDB, KEGG, MetaCyc, and KNApSAcK have been used to produce a pure list of natural products; however, the list was too small to fully annotate metabolite signals (Fig. 2b). To increase the annotation coverage, giant databases such as PubChem and a set of theoretically possible molecular formula have been employed. One of the drawbacks of using large databases is an increased FDR in search results due to chance similarities derived from many useless entries such as those of synthetic compounds. For example, in the automated methods used in CASMI 2013, the search results listed up to 500 structures, including many false positives. It has been reported that in a search involving the Fourier transform (FT)-MS data (error <0.5 mDa) against the PubChem database, more than 99% of the query mass spectra hit two or more molecular formulae, including false positives. This observation suggested that smaller databases that include only natural products might reduce the FDR in metabolite annotation. Whereas chemical ontology based on MeSH (Medical Subject Headings) was introduced, natural products could not be simply distinguished from synthetic drugs in the PubChem entries. To extract natural products from PubChem entries, the nature of the natural products has to be characterized based on the certain structural parameters. Heteroatom (both oxygen and nitrogen) content, as well as the presence of particular ring systems have been used to distinguish between natural and synthetic products.

Another approach that has been employed to reduce the FDR is to perform a combined analysis of two metabolite properties. Chance similarities were reduced when an identical metabolite was found in both the MS/MS spectra database and the metabolite database upon using the MS/MS spectra and the exact mass data, respectively, as queries. MolGEN employed another strategy for the combined analysis of MS/MS spectra and exact mass data (http://molgen.de/). This strategy was based on the fact that a correct molecular formula is compatible with the estimated formulae of all fragment ions and neutral losses in the measured MS/MS spectra. Algorithms that employ the combined strategy produced a good estimation of molecular formula in CASMI 2013.

The comparison between the automatic and human identifications shows that one difference is a consideration of the background information on sample origins. The automated methodologies employed in CASMI 2013 ignored the background information. For example, in challenge 1 the background information provided was—"The compound is a secondary metabolite isolated from Solanaceae plants." The winning team reported that they used this biochemical background information for manual identification by using "Solanaceae" as the query word while searching the Chemical Abstracts Service (CAS) compound database. The background information was intentionally added to the challenges in CASMI 2013 because a similar procedure could be automatically performed using information available for the species–metabolite relationship in KNApSAcK and HMDB. It must be pointed out here that the sample origin is useful information that is always available for every metabolome dataset.

**CONVERSION OF METABOLITE STRUCTURE TO MASS SPECTRA**

Prediction of the artificial MS/MS spectra of peptides is a key technology in proteomics. The relatively simple fragmentation rules applicable to lipids made it possible to construct artificial MS/MS spectral databases, which greatly facilitate lipidomics research. However, the fragmentation of other small molecules is rather complex. Recently, the methods used to predict plausible fragments from a metabolite structure have been improved. For example, MetFrag is now able to obtain candidates from compound libraries based on the m/z ratio of the precursor molecule and the agreement between measured and in silico fragments. MetFusion is an improved compound identification method that involves the combined searching of several resources, including MetFrag and MassBank. The impressive performance of the artificial MS/MS spectra-based methods in CASMI 2013 demonstrated that this might be a promising strategy for the identification of small molecules.

**ONE-TO-ONE CORRESPONDENCE BETWEEN MEASURED DATA AND STRUCTURE**

The identifiers used for metabolite annotation should be standardized for metabolomics data sharing. For example, the rules of the CASMI 2013 contest obligated the use of (standard) InChI or Simplified Molecular-Input Line-Entry System (SMILES) codes for answers of category 2 (Best structure identification). However, InChI list or the SMILES code may not be an ideal way to describe search results.
because these formats are difficult to comprehend. Furthermore, the search result may not be the complete structure of the metabolite as structural isomerism results in poor one-to-one correspondence between the structure of the metabolite and its mass spectra. For instance, essentially identical or very similar mass spectra are produced from structurally similar flavonoids (see PR040122 and PR101019 in MassBank). Similarly, the mass spectra from two different enantiomers of the same amino acid may be identical. Therefore, mass spectra often fail to distinguish between two similar structures; this can be a major limitation in mass spectrometry-based metabolomics. An example of such an ambiguity is the annotation of the metabolite signal for 'tryptophan.' Metabolite identifier systems such as CAS and InChIKey have three IDs for the D- (153–94-6 and QIVBCDIJIAIPQS-SECBINFHSA-N), L- (73–22-3 and QIVBCDIJIIAIPQS-VIF-PVBQESA-N) and racemic forms (54–12-6 and QIVBCDIJIJIAIPQS-UHFFFAOYSA-N) of tryptophan. However, these are unsuitable for accurate annotation of the tryptophan signal. This is because living organisms contain significant amounts of the D-form of amino acids; these are not separated from their corresponding L-forms by conventional chromatography. One method to address the metabolite signal for tryptophan is to introduce a metabolite ontology system to deal with 'tryptophan as an attribute of metabolite.' Databases such as ChiBi, MetaCyc, KEGG, and PubChem are now implementing ontology systems to classify intracellular metabolites. Although these systems have ontology terms for "tryptophan," it is still not distinguished from its racemic forms. For example, the ChEBI ontology defines the ontology relationship of tryptophan as follows: "L-tryptophan (CHEBI:16828) is a tryptophan (CHEBI:27897)." However, this tryptophan (CHEBI:27897) is identical to CAS 54–12-6 (the racemic form).

Vocabulary for the partial annotation of plant secondary metabolites from MS/MS spectra is a lawless area in the field of metabolomics, where data integration has been hampered by the distinct annotation terms used among researchers. For example, a secondary metabolite has been annotated as a natural product-like compound that indicates the ambiguous structure of metabolite. The metabolite annotation list produced by a database search often includes false positives. In the automated methods employed in CASMI 2013, the search result is a list of structures that includes many false positives in addition to the correct search result. The number of annotatable metabolites and the FDR can be increased by employing a loose threshold for searching. To determine a suitable threshold, a method for the elucidation of FDR is required. The reliability of proteomics and transcriptomics research depends on the methods employed to estimate FDR levels. In discovery proteomics, a target-decoy method has been employed for the estimation of FDR levels in the peptide identification list. A decoy database includes the MS/MS spectra of the reversed sequences of all peptides in the real or target database that scarcely overlap each other. A set of acquired MS/MS spectra were searched against both the target and the decoy databases. Since the hits on the decoy database are by chance, the FDR could be defined as the ratio of the number of hits on the decoy database to the number of hits on the target database. Although the target-decoy strategy may be useful in metabolite annotation, methods to build decoy MS/MS spectra and exact mass number of metabolites have been unknown. To construct such a database, a list of natural product-like compounds that can never be detected from real samples has to be prepared. However, as discussed above, the identity of all the metabolites that are present in a sample extract are not known, implying that the target-decoy method might be irrelevant for metabolomics.

BLAST is the de facto method employed for searching amino acid and nucleotide sequence databases.35 In this method, the statistical significance of the similarity between two nucleotide or amino acid sequences are evaluated by Karlin-Altschul statistics as the probability of obtaining the similarity score by chance (p-value).35,36 The FDR can be controlled by specifying a suitable threshold for the search. On the other hand, the de facto standard for searching MS spectral databases (cosine product) is not a probability-based method. Probability-based methods such as X-Rank and EIMS-BLAST for MS spectral database searching have been reported previously.52,53,54 EIMS-BLAST employs Karlin-Altschul statistics by converting electron ionization (EI)-MS spectra into EI-MS spectral sequences. X-Rank computes the probability that a rank from an experimental spectrum matches a rank from a reference library spectrum. However, both of these methods require scoring parameters that depend on the reference databases and the signal intensity information of fragment ions that are unavailable or unreliable for the literature and artificially generated MS/MS spectra. Further development of probability-based methods will consolidate a foundation for the quality control of metabolite annotation. CASMI will play important role in the further development of the infrastructure available for metabolite annotation and hence, might play an important role in shaping metabolomics research.

Acknowledgement

I thank Prof. T. Nishioka (Nara Institute of Science and Technology) for helpful comment to the manuscript.

REFERENCES

1) J. F. Xiao, B. Zhou, H. W. Ressom. Metabolite identification and quantitation in LC-MS/MS-based metabolomics. Trends Analyt. Chem. 32: 1–14, 2012.
2) F. Matsuda, R. Nakabayashi, Y. Sawada, M. Suzuki, M. Y. Hirai, S. Kanaya, K. Saito. Mass spectra-based framework for automated structural elucidation of metabolome data to explore phytochemical diversity. Front. Plant Sci. 2: 40, 2011.
3) S. Neumann, S. Becker. Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules.

© 2014 The Mass Spectrometry Society of Japan
5) D. S. Wishart. Advances in metabolite identification. Bioanalysis 3: 1769–1782, 2011.

4) S. Ma, S. K. Chowdhury. Data acquisition and data mining techniques for metabolite identification using LC coupled to high-resolution MS. Bioanalysis 5: 1285–1297, 2013.

10) E. L. Huttlin, A. D. Hegeman, A. C. Harms, M. R. Sussman. Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. J. Proteome Res. 6: 392–398, 2007.

7) E. R. Mardis. Next-generation sequencing platforms. Curr. Opin. Chem. Biol. 17: 4–11, 2013.

12) H. Choi, A. I. Nesvizhskii. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Biometrics 8: 105, 2007.

5) H. Horai, M. Arita, S. Kanaya. KNApSAcK family database. in Biotechnology in Agriculture and Forestry 57 Plant Biotechnol. 26: 479–483, 2009.

11) J. E. Elias, S. P. Gygi. Target-decoy search strategy for increased statistical concepts in mass spectrometry-based proteomics. J. Proteome Res. 7: 47–50, 2008.

3) RetHInKIng SMALL Molecule IDENTIFIcatIon

3) RetHInKIng SMALL Molecule IDENTIFIcatIon

8) F. Matsuda, Y. Shinbo, A. Oikawa, M. Y. Hirai, O. Fiehn, S. Kanaya, K. Saito. Assessment of metabolome annotation quality: A method for evaluating the false discovery rate of elemental composition searches. PLoS ONE 6: e28966, 2011.

14) H. Horai, M. Arita, S. Kanaya, K. Saito. RIKEN tandem mass spectral database (ReSpect) for metabolite-likeness. BMC Bioinformatics 8: 105, 2007.

19) J. E. Peironcely, T. Reijmers, L. Coulier, A. Bender, T. Hanke-meyer. Understanding and classifying metabolite space and metabolite likeness. PLoS ONE 6: e28966, 2011.

10) T. Kind, M. Scholz, O. Fiehn. How large is the metabolome? A critical analysis of data exchange practices in chemistry. PLoS ONE 4: e5440, 2009.

18) H. Kigawa, M. Arita, K. Saito, D. Ohta, D. Shibata, S. Kanaya. Dissection of genotype–phenotype associations in rice grains using metabolite-likeness. J. Proteome Res. 11: 148, 2012.

28) M. A. Hamdalla, D. W. Mandoiu II, D. W. Hill, S. Rajasekaran, D. F. Grant. BioSM: Metabolomics tool for identifying endogenous mammalian biochemical structures in chemical structure space. J. Chem. Inf. Model. 53: 601–612, 2013.

27) F. Matsuda, H. Redestig, Y. Sawada, Y. Shinbo, M. Y. Hirai, S. Kanaya, K. Saito. Visualization of metabolite identifier information. Plant Biotechnol. 26: 479–483, 2009.

24) T. Kind, O. Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 8: 105, 2007.

26) M. Meringer, S. Reinker, J. Zhang, A. Muller. MS/MS data improves automated determination of molecular formulas by mass spectrometry. MATCH Commun. Math. Comput. Chem. 65: 259–290, 2011.

25) J. E. Peironcely, T. Reijmers, L. Coulier, A. Bender, T. Hanke-meyer. Understanding and classifying metabolite space and metabolite likeness. PLoS ONE 6: e28966, 2011.