Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Advances in computational metabolomics and databases deepen the understanding of metabolisms

Hiroshi Tsugawa[1,2]

Mass spectrometry (MS)-based metabolomics is the popular platform for metabolome analyses. Computational techniques for the processing of MS raw data, for example, feature detection, peak alignment, and the exclusion of false-positive peaks, have been established. The next stage of untargeted metabolomics would be to decipher the mass fragmentation of small molecules for the global identification of human-, animal-, plant-, and microbiota metabolomes, resulting in a deeper understanding of metabolisms. This review is an update on the latest computational metabolomics including known/expected structure databases, chemical ontology classifications, and mass spectrometry cheminformatics for the interpretation of mass fragmentations and for the elucidation of unknown metabolites. The importance of metabolome 'databases' and 'repositories' is also discussed because novel biological discoveries are often attributable to the accumulation of data, to relational databases, and to their statistics. Lastly, a practical guide for metabolite annotations is presented as the summary of this review.

Addresses
[1] RIKEN Center for Sustainable Resource Science, Japan
[2] RIKEN Center for Integrative Medical Sciences, Japan

Corresponding author: Tsugawa, Hiroshi (hiroshi.tsugawa@riken.jp)

Current Opinion in Biotechnology 2018, **54**:10–17

This review comes from a themed issue on **Analytical biotechnology**

Edited by **Hiroshi Shimizu** and **Fumio Matsuda**

https://doi.org/10.1016/j.copbio.2018.01.008

Why is untargeted metabolomics needed in biology?

Under the central dogma, the genome, transcriptome, and proteome are presented in terms of a 'signal flow' and the metabolome is considered the 'result' in metabolism. However, many studies have reported that the metabolites themselves are deeply involved in the physiological functions and homeostasis of living organisms. Examples are first, oxylipins [1], a oxidized fatty acids group that acts as bioactive metabolites in, for example, inflammatory responses and defense systems; second, oncometabolites [2–3], unexpected products from altered metabolism that are involved in tumorigenesis; third, damaged metabolites [4], chemically reactive compounds resulting from enzyme errors or spontaneous reactions that are normally regulated by damage-control systems; fourth, microbiota metabolites [5], metabolites secreted by gut microbiota affecting the host physiology; and finally phytochemicals [6], the plant specialized metabolites exerting various bioactivities on human metabolisms (Figure 1).

Mass spectrometry (MS)-based untargeted metabolomics has led to the discovery of these metabolites and updates on analytical chemistry and its informatics are essential for the elucidation of new physiological function and biological mechanisms.

What is needed to improve untargeted metabolomics?

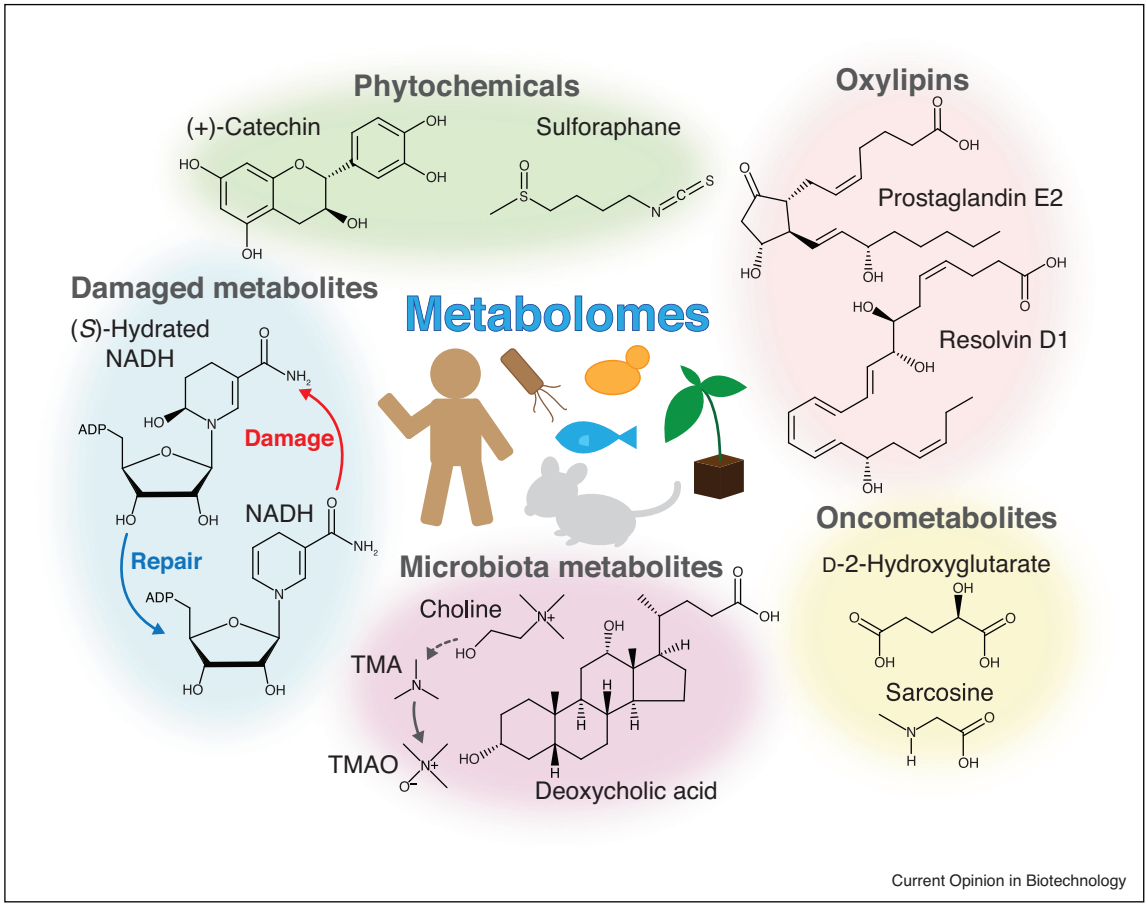
The handling of MS raw data, for example, feature detection, chromatogram deconvolution, isotope recognition, chromatogram alignment, and the exclusion of false-positive peaks is now a mature technique for untargeted metabolomics: of course, the advances also enhance the efficiency for biological discoveries. Software programs such as MS-DIAL [7], MZmine [8], XCMS [9], OpenMS [10], and other specialized programs for metabolomics and lipidomics are used as the pipeline of the metabolomics workflow [11–12]; the favorite program can be used while considering their advantages and disadvantages.

The biggest challenge is the decoding of physics/chemical phenomena of ionized metabolites such as ion interactions [13] (e.g. dimers, adduct ions) and mass fragmentations including in-source fragmentation and low-energy collision-induced dissociation-based fragmentations in mass spectrometers [14•]. Such knowledge will make ion feature detection more efficient and facilitate the global identification of metabolites in living organisms. To date, the 'computational mass fragmentation' using cheminformatics platforms like chemistry development kits [15] are the popular technique to assist the interpretation of mass fragmentations and to elucidate unknown structures with metabolome databases and repositories [16], which is presented below.

Cheminformatics using spectral databases and structure databases

First, the current MS/MS spectral and biologically reported/expected structure databases were examined

Figure 1



Metabolomes linked to physiological functions. The screening of metabolomes is frequently performed by untargeted metabolomics. Bioactive metabolites are validated by targeted analysis for stereoisomer determinations in combination with other analytical platforms such as nucleic magnetic resonance (NMR) and X-ray. The abbreviations TMA and TMAO mean that trimethylamine and trimethylamine *N*-oxide, respectively.

for this review. The statistics was performed by RIKEN internal MS/MS spectral databases including our internal database, MassBank, GNPS, Metlin, ReSpect, and NIST14 (for spectrum count) and the structure databases of MS-FINDER version 2.24 [17••] that include 15 metabolome structure databases (for structure counts). As a result, 226,204 unique compounds were stored in the metabolome structure database whereas the MS/MS spectrum for 7195 compounds of these was recorded in the spectral database, where the first layer of InChIKey was used as the query. Computational metabolomics attempts to fill the large 'gap' between spectrum and structure counts. For a better understanding of the required technologies, the 'metabolome' is divided to four classes in this review, firstly, 'Known Structure-Known Spectrum (KS-KS)' where the reported structure is confirmed by the experimental MS/MS spectrum; secondly, 'Known Structure-Unknown Spectrum (KS-US)' where the biologically examined (or partially expected) structures for which the spectrum is not validated by standard compounds; thirdly, 'Unknown Structure-Known Spectrum (US-KS)' where the mass spectrum itself is frequently monitored in biological samples but the structure is not elucidated or reported in life-science papers; and finally, 'Unknown Structure-Unknown Spectrum (US-US)' where the putative dark matter of small molecules is unknown [18].

The identification of KS-KS metabolites is relatively easy with the aid of EI-MS and MS/MS matching algorithms [19,20•,21,22] combined with retention-time predictions [23,24,25•], and by means of the internal standards. Notably, study-dependent false discovery rate (FDR) estimations have recently been proposed in metabolomics [26••] while a platform-independent annotation rule of lipids has been proposed in lipidomics [27••]; they may facilitate the full automation of the metabolomics/lipidomics workflow.

A challenge in mass spectrometry cheminformatics is the annotation of KS-US and US-KS metabolites, and it has been met by three major computational approaches: the

extrapolation of spectrum knowledge to structurally similar or same scaffold compounds as used in LipidBlast family [28••,29–31], PlantMAT [32•], FlavonoidSearch [33•] (type A); searching for reported molecular structures followed by ranking the structure candidates with the evaluation techniques that untangle structure–spectrum relationships as used in CSI:FingerID [34••], MAGMA [35], MetFrag [36], CFM-ID [37••], MIDAS [38], and MS-FINDER [17••] (type B); and genome scale or molecular spectrum networking approaches to mine the common features of product ions and neutral losses as used in GNPS [39••], MS2LDA [40•], BioCAn [41•], and others [42,43•] (type C). In principle, these programs can be used for the annotation of KS-US and US-KS metabolites; applied in combination, they will contribute to the feature finding of product ions and neutral losses defining specific metabolite class and to the deeper understanding of mass fragmentations.

Notably, type B requires suitable structure databases for searching the chemical spaces. In category 3 of CASMI 2017, all participants used MS-FINDER [17••] for structure assignments in which the team headed by Dr. Tobias Kind outperformed all others (http://www.casmi-contest.org/2017/index.shtml). One of the reasons is that the Kind team carefully optimized the target structure databases; it correctly assigned 37% (91/243), 61% (148/243), and 79% (193/243) challenges as the top, top 3, and top 10 candidates, respectively. This suggests that compound identification can be drastically improved by database selections and curations in specific organs, tissues, and species. Especially in natural product research, taxonomical filters that apply information on species–chemicals relationships efficiently exclude false-positive candidates. In fact, the CASMI contest is very important not only for the activation of computational mass spectrometry but also for the awareness of practically required methods in metabolomics [44•].

Chemical ontologies and the classification system will facilitate metabolite annotations in biology

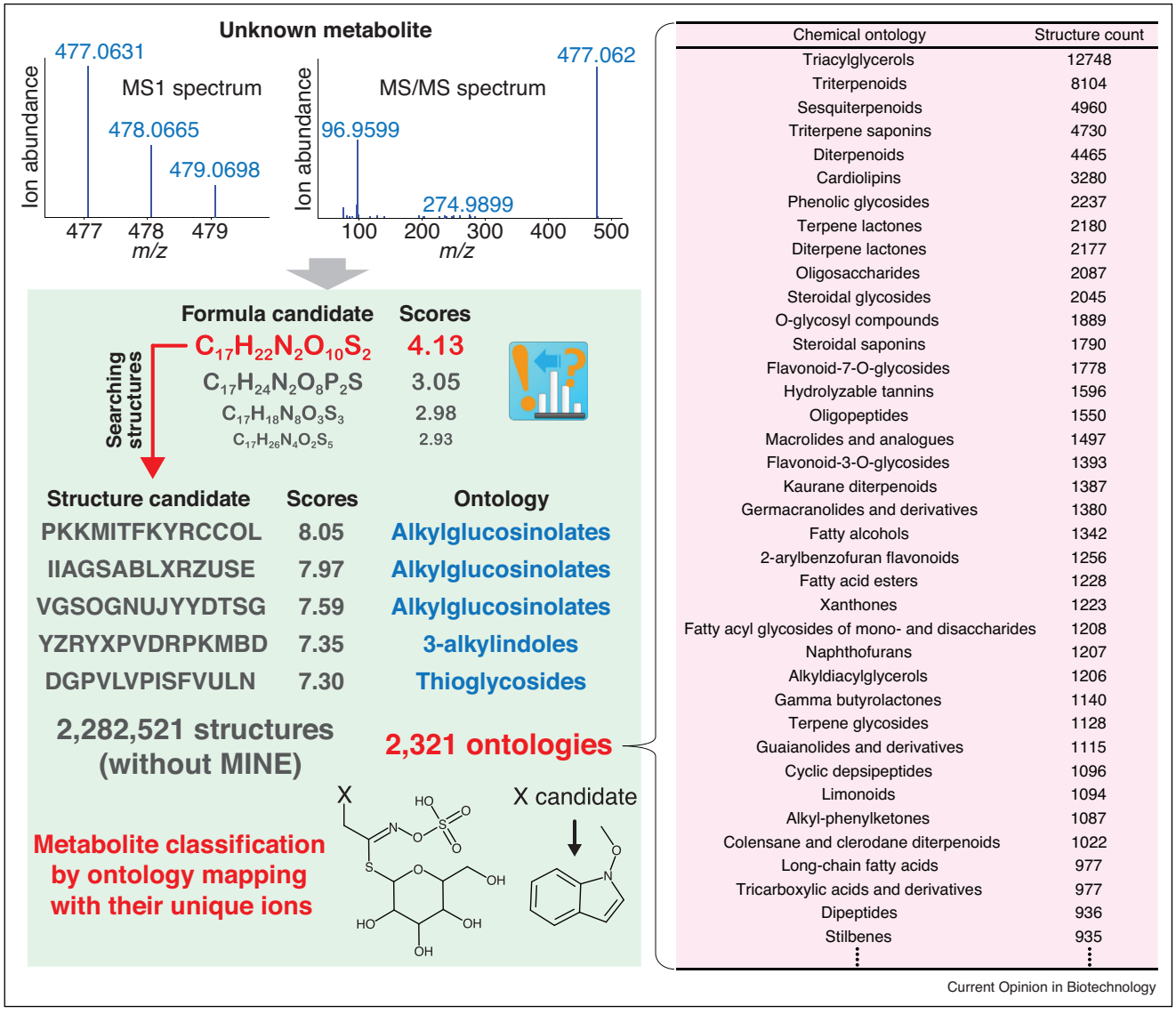
'Metabolite classification' for unknown spectra is an essential technique for structure elucidations. The diversity of small molecules continues to grow; in December 2017, the counts of chemical structures in HMDB [45•], ChemSpider [46], and PubChem [47] compounds are 114,103, >61 million, and >90 million, respectively. As these spaces cannot be comprehended (and most of them cannot be handled by the current metabolomics programs), their condensation into a chemical classification system for the filtering, organizing, and querying of chemicals and for linking to other omics layers as used in multi-omics studies is desirable. Chemical ontology/taxonomy terms have been organized by several teams in MeSH [48], LipidMAPS [49], ChEBI [50], and ClassyFire ChemOnt [51••]; classification can be performed systematically by the related programs [51••,52,53]. Such information on chemical ontologies would also prompt metabolite annotations when using the structure elucidation tools described above.

Figure 2 shows the MS-FINDER result of structure elucidation, querying *m/z* 477.0631 derived from *Arabidopsis thaliana*. While the molecular formula $C_{17}H_{22}N_2O_{10}S_2$ was predicted as the top candidate with a significantly higher score than the others, the determination of structural isomers by the scores is difficult. On the other hand, the substructures of this molecule can be intuitively determined from the assigned ontologies; in the example, the structure may contain the moieties of 'glucosinolates', 'indoles', and 'glycosides'. The ontology terms can also be used for the refinement of molecular-networking approaches [54••,55,56]. In fact, chemical ontology determination for unknown EI-MS or MS/MS spectra is required for the dereplication of natural products [57] and for the exploration of novel compound scaffolds in various species and tissues, including specific plants and microbiomes [58,59••].

The importance of metabolomics databases and repositories

The most challenging issue in mass spectrometry cheminformatics is the elucidation of US-KS or US-US metabolites whose structures are unreported but expected in current biological research. As biology and MS experts succeeded in the identification of unexpected metabolites with a lot of time and effort, the significance, relevance, and occurrence among species, tissues, and organs should be evaluated by investigating metabolomics repositories before annotation. The Metabolomics Workbench [60] and MetaboLight [61] are repositories of MS raw data, and 'in principle', relational searching of such data may shed light on the relevance and occurrence of unknown spectra. On the other hand, these investigations demand MS data integrity, and relational 'databases' for querying the targeted unknown peaks must be developed: this would be a challenging issue of current metabolomics repositories. While the linking of unidentified metabolites is not easy in LC-MS (yet) even by using the retention time, accurate *m/z*, isotopic patterns, and the MS/MS spectrum as the compound property, success in GC–MS-based metabolomics has been documented recently [62••]. The GC–MS BinBase metabolome database associates known and unknown metabolites by the robust retention index, the scalable 70 eV EI-MS spectrum, and other chromatographic properties; the statistics of ion abundances of a specific unknown metabolite can be examined by the BinVestigate web service. The unknowns (actually US-KS metabolites) evaluated as biologically important metabolites by BinVestigate were identified by additional cheminformatics approaches using MS-DIAL [12] and MS-FINDER [17••]. Consequently, metabolomics repositories and the related

Figure 2



A result of MS-FINDER structure elucidation showing the efficiency of chemical ontology assignments. An example for querying *m/z* 477.0631 is shown. The scores for ranking molecular formula and structure candidates are calculated by MS-FINDER version 2.24 which contains a total of 2,282,521 metabolome structures as the search space. The chemical ontologies are defined by 'direct parent' from ClassyFire program, and the structures are currently classified to a total of 2321 chemical ontologies. Right table shows a part of details of an ontology and its structure count included in MS-FINDER.

databases would facilitate the discovery of new metabolites that are not explained by current genome sequences and known metabolic pathways.

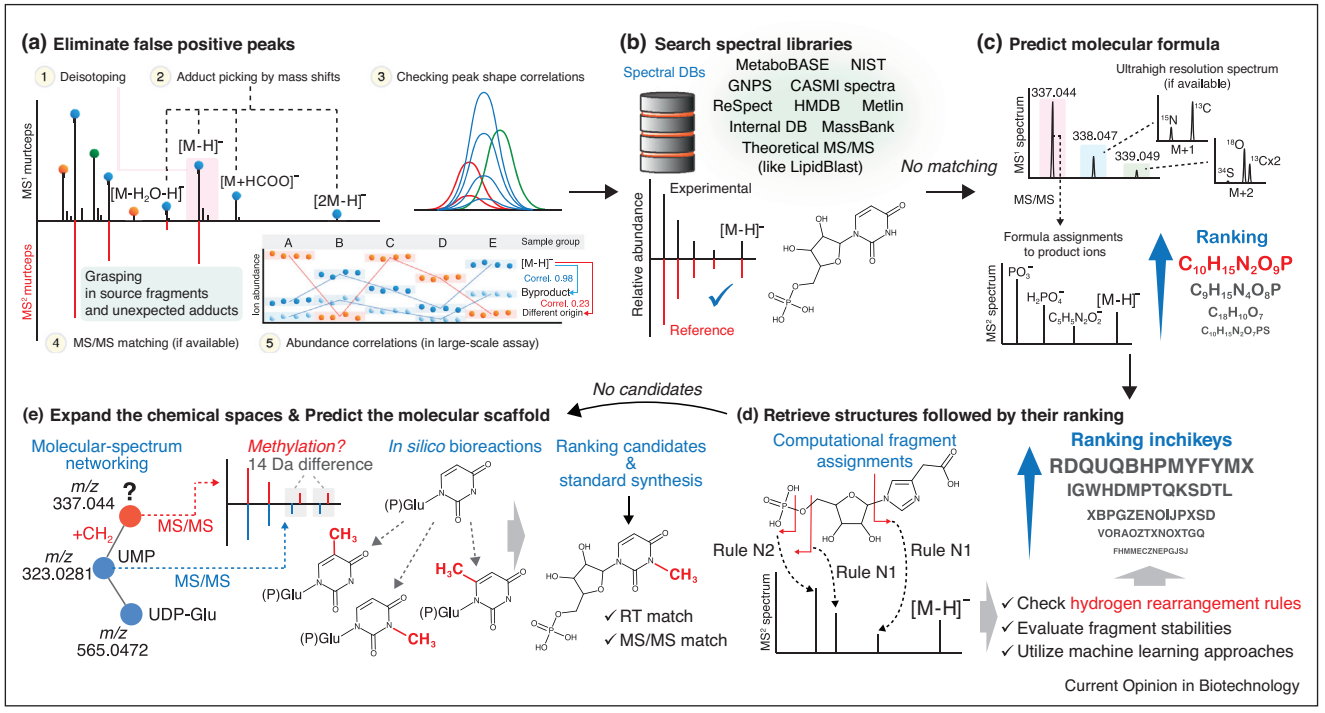
The 'guide' for metabolite annotation by current MS-based cheminformatics

Mass spectrometry cheminformatics would expand the coverage of metabolite identification and annotation in untargeted metabolomics. The signpost for metabolite discoveries is shown as the summary of this review (Figure 3).

Workflow:

1. *Eliminate the possibility of false-positive peaks:* Although this is not the focus of this review, false-positive peaks and their spectra thought to be isotopic ions, different adduct types, in-source fragments, and other background ions should be excluded before annotation [63,64]. Curation can be assisted by several programs such as CAMERA [65], MS-FLO [66], RAMClust [67], xMSannotator [68], and the internal functions of metabolomics software programs. In addition to the

Figure 3



The practical workflow for metabolite annotation. **(a)** False-positive peaks are excluded by deisotoping, adduct picking, peak grouping by peak shapes, MS/MS matching to the MS1 spectrum, and by ion abundance correlations among sample groups. **(b)** Mass spectral databases are used first. **(c)** and **(d)** If no spectrum is available, molecular formula prediction followed by structure elucidation are performed. **(e)** If no candidate is retrieved, the molecular scaffold and modifications from known chemicals can be predicted with molecular-spectrum networking and chemical space expansion in combination with structure-elucidation programs.

classical adduct/in-source fragment detection methods based on the correlations of chromatogram data points, recent programs apply MS/MS matching to the MS1 spectrum and the correlation factor of ion abundances among analyzed sample groups to suggest in-source fragments and unexpected adduct ions. Hopefully, purification and concentration of unknown metabolites are required to increase the ion abundances of the mass spectrum.

2. *Search spectral libraries:* The first choice for structure elucidation is mass spectral searching with publicly and commercially available spectral databases. In addition to the normal use for spectral searching of the tandem mass (MS/MS) spectrum, the search space can be expanded to all records by not using precursor isolation because product ion similarities often provide direct evidence for the substructures and molecular scaffolds of unknown metabolites (see below).
3. *Predict the molecular formula:* The first task for unknown molecules in MS is to determine the molecular formula. Programs like MolecularWeightCalculator (https://omics.pnl.gov/software/molecular-weight-calculator), Sirius [69], and MS-FINDER [17••] with seven golden rules [70] assist prediction, and ultra-high resolution MS can provide exact oxygen, nitrogen, and sulfur counts of the molecular formula [71]. Furthermore, labeling approaches using the fully labeled samples by ^{13}C, ^{15}N, ^{18}O, or ^{34}S chemicals can be used for the strict determination of formula element count for unknown metabolite [72,73].
4. *Retrieve known/expected structures of a suggested formula, followed by their ranking:* That most unknowns can be contained in metabolome structure databases is a working hypothesis. There are several cheminformatics programs for searching databases followed by ranking the structures as introduced in this review. If the formula is found in databases, the top 10 structural candidates are the practical targets. Additional necessary criteria including retention time/index predictions and taxonomical information on targeted species can be obtained from several platforms such as PredRet [25•] and NIST RI [74] (for retention time prediction) and from databases such as HMDB [45•] and KNApSAcK [75•] (for taxonomical information).
5. *Expand the chemical spaces for searching and predict the molecular scaffold*: If there is no information for structure in databases, structure elucidation is very difficult. The computationally expanded chemical spaces obtained with biologically expected chemical reactions in, for example, MINE [76] and LipidHome [77] are useful. Molecular-spectrum networking

[54••] also helps to elucidate the scaffold by extracting the common features of product ion or neutral losses with the known spectrum of the chemical. In addition, chemical classifications utilizing mass spectrum features assist in compound annotation [59••].

Additional approaches using genome-scale information [41•], bioreaction knowledge [42], ion abundance correlation networks [68], and accumulated metabolomics databases/repositories [62••] are also incorporated. Overall, the cheminformatics techniques that were developed in drug discovery research are now widely utilized in MS-based metabolomics studies. Technological advances in mass spectrometry informatics as well as bioinformatics for the interpretation of metabolome data deepen the understanding of metabolisms.

Acknowledgement

This review was written with the support of JSPS KAKENHI Grant Number 15K01812.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Serhan CN: **Treating inflammation and infection in the 21st century: new hints from decoding resolution mediators and mechanisms**. *FASEB J* 2017, **31**:1273-1288.
2. Xu W, Yang H, Liu Y, Yang Y, Wang P, Kim SH, Ito S, Yang C, Wang P, Xiao MT *et al.*: **Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases**. *Cancer Cell* 2011, **19**:17-30.
3. Locasale JW: **Serine, glycine and one-carbon units: cancer metabolism in full circle**. *Nat Rev Cancer* 2013, **13**:572-583.
4. Linster CL, Van Schaftingen E, Hanson AD: **Metabolite damage and its repair or pre-emption**. *Nat Chem Biol* 2013, **9**:72-80.
5. Rooks MG, Garrett WS: **Gut microbiota, metabolites and host immunity**. *Nat Rev Immunol* 2016, **16**:341-352.
6. Wang H, Oo Khor T, Shu L, Su Z-Y, Fuentes F, Lee J-H, Tony Kong A-N: **Plants against cancer: a review on natural phytochemicals in preventing and treating cancers and their druggability**. *Anticancer Agents Med Chem* 2012, **12**:1281-1305.
7. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, Kanazawa M, VanderGheynst J, Fiehn O, Arita M: **MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis**. *Nat Methods* 2015, **12**:523-526.
8. Pluskal T, Castillo S, Villar-Briones A, Oresic M: **MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data**. *BMC Bioinformatics* 2010, **11**:395.
9. Mahieu NG, Genenbacher JL, Patti GJ: **A roadmap for the XCMS family of software solutions in metabolomics**. *Curr Opin Chem Biol* 2016, **30**:87-93.
10. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich H-C, Gutenbrunner P, Kenar E *et al.*: **OpenMS: a flexible open-source software platform for mass spectrometry data analysis**. *Nat Meth* 2016, **13**:741-748.
11. Spicer R, Salek RM, Moreno P, Cañueto D, Steinbeck C: **Navigating freely-available software tools for metabolomics analysis**. *Metabolomics* 2017, **13**:1-16.
12. Tsugawa H, Ikeda K, Arita M: **The importance of bioinformatics for connecting data-driven lipidomics and biological insights**. *Biochim Biophys Acta Mol Cell Biol Lipids* 2017, **1862**:762-765.
13. Kostiainen R, Kauppila TJ: **Effect of eluent on the ionization process in liquid chromatography–mass spectrometry**. *J Chromatogr A* 2009, **1216**:685-699.
14. • Demarque DP, Crotti AEM, Vessecchi R, Lopes JLC, Lopes NP: **Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products**. *Nat Prod Rep* 2016, **33**:432-455.
15. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E: **The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics**. *J Chem Inf Comput Sci* 2003, **43**:493-500.
16. Scheubert K, Hufsky F, Böcker S: **Computational mass spectrometry for small molecules**. *J Cheminform* 2013, **5**:1-24.
17. •• Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, Saito K, Fiehn O, Arita M: **Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software**. *Anal Chem* 2016, **88**:7946-7958.
18. Showalter MR, Cajka T, Fiehn O: **Epimetabolites: discovering metabolism beyond building and burning**. *Curr Opin Chem Biol* 2017, **36**:70-76.
19. Stein S: **Mass spectral reference libraries: an ever-expanding resource for chemical identification**. *Anal Chem* 2012, **84**:7274-7282.
20. • Matsuda F, Tsugawa H, Fukusaki E: **Method for assessing the statistical significance of mass spectral similarities using basic local alignment search tool statistics**. *Anal Chem* 2013, **85**:8291-8297.
21. Mylonas R, Mauron Y, Masselot A, Binz P-A, Budin N, Fathi M, Viette V, Hochstrasser DF, Lisacek F: **X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry**. *Anal Chem* 2009, **81**:7604-7610.
22. Kind T, Tsugawa H, Cajka T, Ma Y, Lai Z, Mehta SS, Wohlgemuth G, Barupal DK, Showalter MR, Arita M *et al.*: **Identification of small molecules using accurate mass MS/MS search**. *Mass Spectrom Rev* 2017 http://dx.doi.org/10.1002/mas.21535.
23. Matsuo T, Tsugawa H, Miyagawa H, Fukusaki E: **Integrated strategy for unknown EI-MS identification using quality control calibration curve, multivariate analysis, EI-MS spectral database, and retention index prediction**. *Anal Chem* 2017, **89**:6766-6773.
24. Cao M, Fraser K, Huege J, Featonby T, Rasmussen S, Jones C: **Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics**. *Metabolomics* 2015, **11**:696-706.
25. • Stanstrup J, Neumann S, Vrhovek U: **PredRet: prediction of retention time by direct mapping between multiple chromatographic systems**. *Anal Chem* 2015, **87**:9421-9428.
26. •• Scheubert K, Hufsky F, Petras D, Wang M, Nothias L-F, Dührkop K, Bandeira N, Dorrestein PC, Böcker S: **Significance estimation for large scale metabolomics annotations by spectral matching**. *Nat Commun* 2017, **8**:1494.
27. •• Hartler J, Triebl A, Ziegl A, Trötzmüller M, Rechberger GN, Zeleznik OA, Zierler KA, Torta F, Cazenave-Gassiot A, Wenk MR *et al.*: **Deciphering lipid structures based on platform-independent decision rules**. *Nat Methods* 2017, **14**.
28. •• Kind T, Liu K-H, Lee DY, DeFelice B, Meissen JK, Fiehn O: **LipidBlast in silico tandem mass spectrometry database for lipid identification**. *Nat Methods* 2013, **10**:755-758.
29. Kind T, Okazaki Y, Saito K, Fiehn O: **LipidBlast templates as flexible tools for creating new in-silico tandem mass spectral libraries**. *Anal Chem* 2014, **86**:11024-11027.

30. Ma Y, Kind T, Vaniya A, Gennity I, Fahrmann JF, Fiehn O: **An in silico MS/MS library for automatic annotation of novel FAHFA lipids**. *J Cheminform* 2015, **7**:2-6.

31. Tsugawa H, Ikeda K, Tanaka W, Senoo Y, Arita M, Arita M: **Comprehensive identification of sphingolipid species by in silico retention time and tandem mass spectral library**. *J Cheminform* 2017, **9**:1-12.

32. • Qiu F, Fine DD, Wherritt DJ, Lei Z, Sumner LW: **PlantMAT: a metabolomics tool for predicting the specialized metabolic potential of a system and for large-scale metabolite identifications**. *Anal Chem* 2016, **88**:11373-11383.

33. • Akimoto N, Ara T, Nakajima D, Suda K, Ikeda C, Takahashi S, Muneto R, Yamada M, Suzuki H, Shibata D *et al.*: **FlavonoidSearch: a system for comprehensive flavonoid annotation by mass spectrometry**. *Sci Rep* 2017, **7**:1-9.

34. •• Dührkop K, Shen H, Meusel M, Rousu J, Böcker S: **Searching molecular structure databases with tandem mass spectra using CSI:FingerID**. *Proc Natl Acad Sci U S A* 2015 http://dx.doi.org/10.1073/pnas.1509788112.

35. Ridder L, van der Hooft JJJ, Verhoeven S: **Automatic compound annotation from mass spectrometry data using MAGMa**. *Mass Spectrom* 2014, **3** S0033-S0033.

36. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S: **MetFrag relaunched: incorporating strategies beyond in silico fragmentation**. *J Cheminform* 2016, **8**:1-16.

37. •• Allen F, Greiner R, Wishart D: **Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification**. *Metabolomics* 2014, **11**:98-110.

38. Wang Y, Kora G, Bowen BP, Pan C: *MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics*. 2014.

39. •• Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T *et al.*: **Sharing and community curation of mass spectrometry data with global natural products social molecular networking**. *Nat Biotechnol* 2016, **34**:828-837.

40. • van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S: **Topic modeling for untargeted substructure exploration in metabolomics**. *Proc Natl Acad Sci U S A* 2016, **113**:13738-13743.

41. • Alden N, Krishnan S, Porokhin V, Raju R, McElearney K, Gilbert A, Lee K: **Biologically consistent annotation of metabolomics data**. *Anal Chem* 2017 http://dx.doi.org/10.1021/acs.analchem.7b02162.

42. Morreel K, Saeys Y, Dima O, Lu F, Van de Peer Y, Vanholme R, Ralph J, Vanholme B, Boerjan W: **Systematic structural characterization of metabolites in *Arabidopsis* via candidate substrate-product pair networks**. *Plant Cell* 2014, **26**:929-945.

43. • Li D, Heiling S, Baldwin IT, Gaquerel E: **Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory**. *Proc Natl Acad Sci U S A* 2016, **113**:E7610-E7618.

44. • Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, Allen F, Vaniya A, Verdegem D, Böcker S *et al.*: **Critical assessment of small molecule identification 2016: automated methods**. *J Cheminform* 2017, **9**:1-21.

45. • Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, Sajed T, Johnson D, Li C, Karu N *et al.*: **HMDB 4.0: the human metabolome database for 2018**. *Nucleic Acids Res* 2017 http://dx.doi.org/10.1093/nar/gkx1089.

46. Pence HE, Williams A: **Chemspider: an online chemical information resource**. *J Chem Educ* 2010, **87**:1123-1124.

47. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA *et al.*: **PubChem substance and compound databases**. *Nucleic Acids Res* 2016, **44**:D1202-D1213.

48. Leydesdorff L, Opthof T: **Citation analysis with medical subject headings (MeSH) using the Web of Knowledge: a new routine**. *J Assoc Inf Sci Technol* 2013, **64**:1076-1080.

49. Fahy E, Sud M, Cotter D, Subramaniam S: **LIPID MAPS online tools for lipid research**. *Nucleic Acids Res* 2007, **35**:606-612.

50. Degtyarenko K, De matos P, Ennis M, Hastings J, Zbinden M, Mcnaught A, Alcántara R, Darsow M, Guedj M, Ashburner M: **ChEBI: a database and ontology for chemical entities of biological interest**. *Nucleic Acids Res* 2008, **36**:344-350.

51. •• Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, Fahy E, Steinbeck C, Subramanian S, Bolton E *et al.*: **ClassyFire: automated chemical classification with a comprehensive, computable taxonomy**. *J Cheminform* 2016, **8**:1-20.

52. Bobach C, Böhme T, Laube U, Püschel A, Weber L: **Automated compound classification using a chemical ontology**. *J Cheminform* 2012, **4**:1-12.

53. Hastings J, Magka D, Batchelor C, Duan L, Stevens R, Ennis M, Steinbeck C: **Structure-based classification and ontology in chemistry**. *J Cheminform* 2012, **4**:8.

54. •• Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM *et al.*: **Mass spectral molecular networking of living microbial colonies**. *Proc Natl Acad Sci U S A* 2012, **109**:E1743-E1752.

55. Nguyen DD, Wu C-H, Moree WJ, Lamsa A, Medema MH, Zhao X, Gavilan RG, Aparicio M, Atencio L, Jackson C *et al.*: **MS/MS networking guided analysis of molecule and gene cluster families**. *Proc Natl Acad Sci U S A* 2013, **110**:E2611-E2620.

56. Grapov D, Wanichthanarak K, Fiehn O: **MetaMapR: pathway independent metabolomic network analysis incorporating unknowns**. *Bioinformatics* 2015, **31**:2757-2760.

57. Kind T, Fiehn O: **Strategies for dereplication of natural compounds using high-resolution tandem mass spectrometry**. *Phytochem Lett* 2017, **21**:313-319.

58. Tsugawa H, Tsujimoto Y, Arita M, Bamba T, Fukusaki E: **GC/MS based metabolomics: development of a data mining system for metabolite identification by using soft independent modeling of class analogy (SIMCA)**. *BMC Bioinformatics* 2011, **12**:131.

59. •• Treutler H, Tsugawa H, Porzel A, Gorzolka K, Tissier A, Neumann S, Balcke GU: **Discovering regulated metabolite families in untargeted metabolomics studies**. *Anal Chem* 2016, **88**:8082-8090.

60. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS *et al.*: **Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools**. *Nucleic Acids Res* 2016, **44**:D463-D470.

61. Kale NS, Haug K, Conesa P, Jayseelan K, Moreno P, Rocca-Serra P, Nainala VC, Spicer RA, Williams M, Li X *et al.*: **MetaboLights: an open-access database repository for metabolomics data**. *Curr Protoc Bioinforma* 2016:14. 13.1–14.13.18.

62. •• Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y, Ogiwara A, Meissen J, Showalter M, Takeuchi K *et al.*: **Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics**. *Nat Methods* 2017 http://dx.doi.org/10.1038/nmeth.4512.

63. Xu YF, Lu W, Rabinowitz JD: **Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography–mass spectrometry-based metabolomics**. *Anal Chem* 2015, **87**:2273-2281.

64. Mahieu NG, Patti GJ: **Systems-level annotation of a metabolomics data set reduces 25 000 features to fewer than 1000 unique metabolites**. *Anal Chem* 2017, **89**:10397-10406.

65. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S: **CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets**. *Anal Chem* 2012, **84**:283-289.

66. DeFelice BC, Mehta SS, Samra S, Čajka T, Wancewicz B, Fahrmann JF, Fiehn O: **Mass spectral feature list optimizer (MS-**

FLO): a tool to minimize false positive peak reports in untargeted liquid chromatography–mass spectroscopy (LC-MS) data processing. *Anal Chem* 2017, **89**:3250-3255.

67. Broeckling CD, Afsar FA, Neumann S, Ben-Hur A, Prenni JE: **RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data**. *Anal Chem* 2014, **86**:6812-6817.

68. Uppal K, Walker DI, Jones DP: **xMSannotator: an R package for network-based annotation of high-resolution metabolomics data**. *Anal Chem* 2017, **89**:1063-1067.

69. Böcker S, Dührkop K: **Fragmentation trees reloaded**. *J Cheminform* 2016, **8**:1-26.

70. Kind T, Fiehn O: **Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry**. *BMC Bioinformatics* 2007, **8**:105.

71. Nakabayashi R, Saito K: **Ultrahigh resolution metabolomics for S-containing metabolites**. *Curr Opin Biotechnol* 2017, **43**:8-16.

72. Nakabayashi R, Sawada Y, Yamada Y, Suzuki M, Hirai MY, Sakurai T, Saito K: **Combination of liquid chromatography-Fourier transform ion cyclotron resonance-mass spectrometry with ^{13}C-labeling for chemical assignment of sulfur-containing metabolites in onion bulbs**. *Anal Chem* 2013, **85**:1310-1315.

73. Nakabayashi R, Hashimoto K, Toyooka K, Saito K: **Top-down metabolomic approaches for nitrogen-containing metabolites**. *Anal Chem* 2017, **89**:2698-2703.

74. Stein SE, Babushok VI, Brown RL, Linstrom PJ: **Estimation of kovats retention indices using group contributions**. *J Chem Inf Model* 2007, **47**:975-980.

75. • Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK *et al.*: **KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research**. *Plant Cell Physiol* 2012, **53**:1-12.

76. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, Hanson AD, Fiehn O, Tyo KEJ, Henry CS: **MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics**. *J Cheminform* 2015, **7**:44.

77. Foster JM, Moreno P, Fabregat A, Hermjakob H, Steinbeck C, Apweiler R, Wakelam MJO, Vizcaíno JA: **LipidHome: a database of theoretical lipids optimized for high throughput mass spectrometry lipidomics**. *PLoS ONE* 2013, **8**:1-8.