Plant Biotechnology 25, 377–380 (2008)

Bioinformatics Note

A tool for high-throughput prediction of molecular formulas and identification of isotopic peaks from large-scale mass spectrometry data

Yukiko Nakamura[1,2], Shigehiko Kanaya[3], Nozomu Sakurai[1], Yoko Iijima[1], Koh Aoki[1], Koei Okazaki[1], Hideyuki Suzuki[1], Masahiko Kitayama[2], Daisuke Shibata[1,*]

[1] Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan; [2] Ehime Women's College, Ibuki, Uwajima, Ehime 798-0025, Japan; [3] Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan

* E-mail: shibata@kazusa.or.jp Tel: +81-438-52-3947 Fax: +81-438-52-3948

Received March 21, 2008; accepted April 23, 2008 (Edited by M. Sekine)

Abstract Most plant metabolites are uncharacterized, even in well-analyzed plant species. High-accuracy measurement of mass values by state-of-the-art mass spectrometers such as Fourier transform ion cyclotron resonance mass spectrometers allows prediction of possible molecular formulas for each metabolite. As a first step in comprehensive metabolite identification from mass spectrometry data, here we have developed a computational tool for high-throughput prediction of molecular formulas and identification of isotopic peaks. The program generates all possible formulas for each mass value under given parameters. To reduce calculation time, monophosphate, diphosphate, triphosphate and bisulfate groups are regarded as monovalent units during formula generation. Prediction of isotopic peaks associated with each metabolite also facilitates reduction in the number of possible formulas. The tool implements these procedures for all mass values from a set of mass spectrometry data. The tool facilitates subsequent annotation of metabolites, which can be integrated with metabolome databases.

Key words: Accurate mass, FTICR-MS, isotopic peaks, mass spectrometer, metabolomics.

Metabolomics is one of the emerging fields of functional genomics (Saito et al. 2008). Plant metabolites number more than 200,000, and most are uncharacterized (Pichersky and Gang 2000; Fiehn 2002; Dixon and Strack 2003). Mass spectrometry is powerful for measuring the mass-to-charge ratio (*m*/*z*) of metabolite-derived ions, and thus has been used for comprehensive quantitative profiling of metabolites and for qualitative elucidation of their chemical structures (Kind and Fiehn 2006). Prediction of the molecular formula of each metabolite peak detected by mass spectrometry is a step in structural elucidation of metabolites. High-accuracy measurement of *m*/*z*, which can be achieved by Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS), is particularly crucial for prediction of molecular formulas (Aharoni et al. 2002; Oikawa et al. 2006; Nakamura et al. 2007; Iijima et al. 2008; Suzuki et al. 2008).

To predict molecular formulas that match observed *m*/*z* values, several calculation tools have been developed. The "Formula To Mass To Formula" (http://www.ch.ic.ac.uk/java/applets/f2m2f/) java applet calculates up to six molecular formulas for which mass values are close to a given mass. The computational tools "Formula Finder" (http://www.alchemistmatt.com/), which is included in the "Molecular Weight Calculator MWTWIN" software package, and "HiRes MS" (http://jhau.maliwi.de/sci/) generate possible molecular formulas with accurate mass and the error from the given mass under given calculation parameters such as the numbers of atoms and mass range. "Formula Predictor" software (Inohara et al. 2006) and the "Formula Generator" of "ACD/MS Manager" software (Advanced Chemistry Development Inc., http://www.acdlabs.com/, Toronto, Ontario, Canada) display molecular formulas predicted from mass values corresponding to individual metabolite peaks under given calculation parameters, which can be set for kinds of atoms, the maximum number of each atom in a single metabolite, error margin of mass values, possible adduct ion forms, double-bond equivalents, hydrogen/carbon ratio and the choice of nitrogen rule. The relative intensity of isotopic ions provides additional information in determining the numbers of atoms (Kind and Fiehn 2006; Iijima et al.

Abbreviations: LC, liquid chromatography; FTICR, Fourier transform ion cyclotron resonance; MS, mass spectrometer.
This article can be found at http://www.jspcmb.jp/

2008). For example, the tools "qmass" (Rockwood et al. 2004) and "emass" (Rockwood and Haimi 2006) calculate masses and intensities of isotopic peaks of a given molecular formula. "Isopro" (http://members.aol.com/msmssoft/) is an isotopic distribution simulator, which was created to look at large biomolecules (Rockwood et al. 1995). These tools are used to execute one-by-one analysis of each metabolite peak. Thus, they are not suitable for large-scale formula prediction of metabolites detected by MS.

Here, we have developed a computational tool for automated, high-throughput prediction of molecular formulas. As the program is written in Java and all inputs and outputs are processed as text files, the tool can be used on common computer platforms.

The computational tool developed in this study comprises three steps: i) generation of molecular formulas for multiple *m*/*z* queries, ii) filtering out the generated molecular formulas by criteria with respect to valence rules, and iii) identification of isotopic peaks associated with individual metabolite mass peaks, which facilitates constraint of molecular formulas. The flow of the program is shown in Figure 1.

First, all possible molecular formulas corresponding to each *m*/*z* value in a mass data file are generated based on accurate masses of atoms. Prior to calculation, the following parameters are set by the user: (1) mass tolerance, (2) the type of adduct ion (choices are "Actual" (not adduct), "$[Actual+NH_4]^+$", "$[Actual+K]^+$", "$[Actual+Na]^+$", "$[Actual+H]^+$", "$[Actual+HCOO]^-$" and "$[Actual+H]^-$"), and (3) the elements to be included (C, H, N, O, S, F, Cl, Br and I) (Figure 2). A novel function of this tool is that functional groups, such as the bisulfate group (HSO_4), monophosphate group (H_2PO_4), diphosphate group ($H_3P_2O_7$) and triphosphate group ($H_4P_3O_{10}$) are taken as monovalent units in the generation of molecular formulas. This efficient approach reduces both the number of generated formulas and the time for calculation. For example, the number of generated formulas including bisulfate was half of the number including sulfur (S), and the time for calculation was also reduced to one-half (Table 1).

The second step constrains candidate molecular

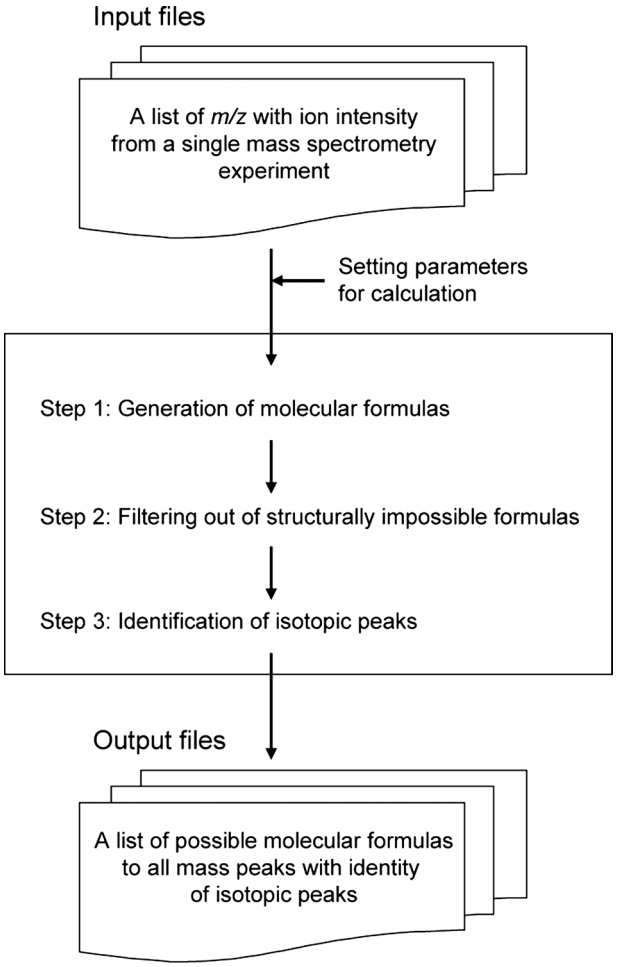

Figure 1. Flowchart of the procedure for possible molecular formula generation and identification of isotopic peaks from a list of mass analysis data.

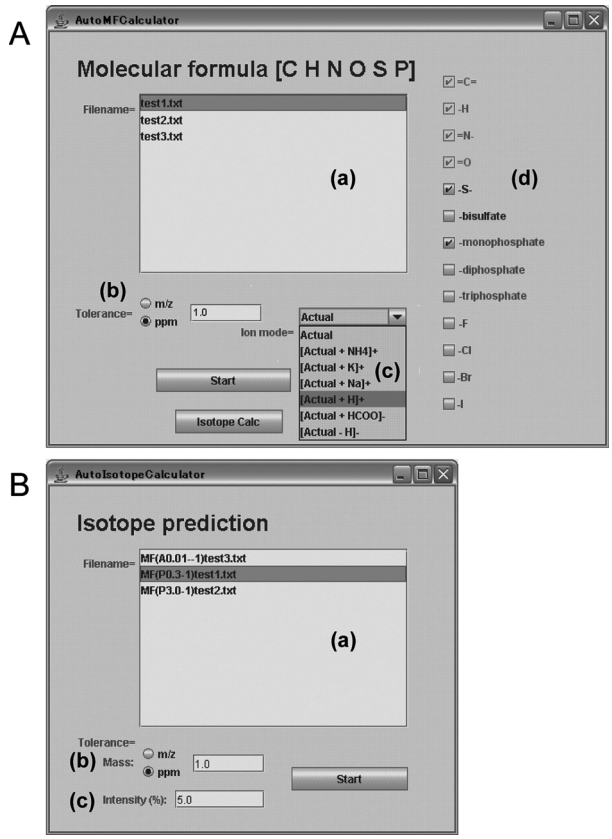

Figure 2. Screenshots illustrating parameter setting and calculation. For molecular formula generation (A), data file(s) should be selected in the filename list (a); calculation parameters such as mass tolerance (b), adduct ion condition (c) and kinds of elemets (d) are selectable. For isotopic peak prediction (B), data file(s) should be selected in the filename list (a); calculation parameters such as tolerance of isotopic mass difference (b) and of isotopic abundance (c) are selectable.

formulas by two criteria. The molecular formula generation described above inevitably produces structurally impossible formulas. Such false-positives exponentially increase as the molecular mass increases, especially in the mass range greater than 500 Da (Kind and Fiehn 2006). Thus, appropriate molecular formulas are filtered out by two criteria: (1) the sum of the valences is an even number, and (2) the sum of the valences of polyvalent elements equals or is more than the sum of the valences of monovalent elements. The latter is equivalent to the criterion that the sum of the valences is greater than or equal to twice the number of atoms minus 1, which was introduced by Kind and Fiehn (2007) as the extended Senior's theorem (Senior 1951).

In the third step, isotopic peaks associated with parental metabolite peaks are identified so the user can reduce the possible molecular formulas for individual metabolites. In a non-targeted analysis, isotopic peaks of most metabolites are under the detection limit of the MS. However, abundant metabolites are, in most cases, associated with isotopic peaks in the mass spectrum. Most of the isotopic peaks detected are derived from single isotope exchanges. Information about isotopic peaks is useful to constrain molecular formulas, especially those of high-mass metabolites >500 Da (Kind and Fiehn 2006). In the first part of the step, the theoretical *m*/*z* value and relative intensity of isotopic signals of the predicted molecular formulas described above are simulated using the polynomial method of Stoll et al. (2006). In the second part, the simulated results for the molecular formulas, which are calculated from *m*/*z* values of observed peak signals as parental metabolite peaks, are compared with other individual peak signals to identify isotopic peaks. If the observed mass difference and the relative intensity between two peaks match theoretical values of any of the simulation patterns, the peaks are regarded as a pair of isotopic peaks of an identical metabolite. Accurate masses and relative abundances of all stable isotopic atoms are calculated according to an IUPAC Technical Report (De Laeter et al. 2003). For example, the procedure for Peak 1 (*m*/*z*=1033.547189) and Peak 2 (*m*/*z*=1034.551582) is shown in Figure 3. The *m*/*z* values output from the MS are represented without consideration for significant figures. The *m*/*z* value of isotopic compounds with single exchanges of ^{12}C to ^{13}C for Candidate 5 of Peak 1 matches the *m*/*z* value of Peak 2. The relative intensity of the isotopic compounds with single exchanges of ^{12}C (natural abundance: 0.9893) to ^{13}C (natural abundance:

Table 1. Number of candidate molecular formulas generated using elements C, H, N, O and S, and using C, H, N, O and bisulfate (HSO_4), and calculation time.

Elements	Candidate formulas	Calculation time (s)
C, H, N, O, S	421	4.125
C, H, N, O, bisulfate (HSO_4)	173	2.281

The formulas were generated from a sample mass value (1254.6291) with ±5 ppm mass tolerance.

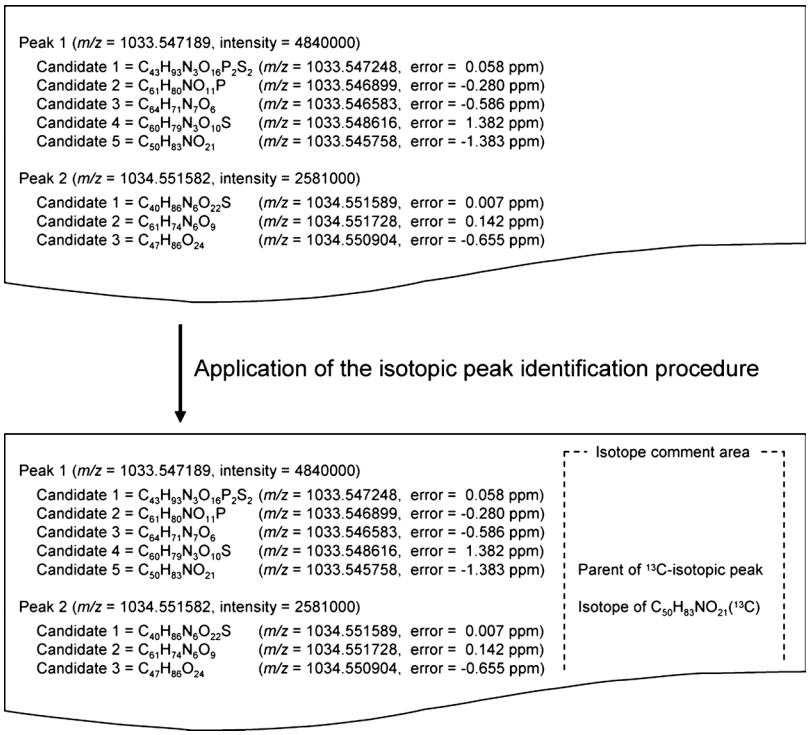

Figure 3. Identification of isotopic peaks. The isotopic peak identification procedure is applied to the list resulting from generating the possible molecular formulas that fit the FTICR-MS data; comments on isotope identification are then added to the output file.

0.0107) for Candidate 5 of Peak 1 (with 50 carbons in the molecular formula) is calculated by the polynomial method:

$0.9893^{49} \times 0.0107^{1} \times 50 = 0.54$

As the intensity of Peak 2 (2581000) is 54% that of Peak 1 (4840000) within ±5% error, it is likely that Candidate 5 for Peak 1 is the parental metabolite of Peak 2. Comments on the isotopic relationship of the peaks are included in the final output list.

The user may include information on isotopes of some metabolites to reduce the number of molecular formulas. For example, a metabolite of 1034.554 $[M+H]^+$, which was detected from LC-FTICR-MS analysis of tomato fruit (Iijima et al. 2008), has 276, 142, 53 or 27 possible molecular formulas, respectively, when the mass tolerance is set to 10 ppm, 5 ppm, 2 ppm or 1 ppm. With information on isotopic peaks, the possible molecular formulas are reduced to 14, 14, 8 and 5, respectively.

In conclusion, the tool presented here provides high-throughput prediction of molecular formulas for a large-scale mass data set. It further accelerates the provision of information obtained by database and reference searches to the mass peaks in metabolite annotation procedures (Iijima et al. 2008). The tool is available on request.

References

Aharoni A, Ric de Vos CH, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R, Goodenowe DB (2002) Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *OMICS* 6: 217–234

De Laeter JR, Bohlke JK, De Bievre P, Hidaka H, Peiser HS, Rosman KJR, Taylor PDP (2003) Atomic weights of the elements: Review 2000 (IUPAC technical report). *Pure Appl Chem* 75: 683–800

Dixon RA, Strack D (2003) Phytochemistry meets genome analysis, and beyond. *Phytochemistry* 62: 815–816

Fiehn O (2002) Metabolomics—the link between genotypes and phenotypes. *Plant Mol Biol* 48: 155–171

Iijima Y, Nakamura Y, Ogata Y, Tanaka K, Sakurai N, Suda K, Suzuki T, Suzuki H, Okazaki K, Kitayama M, Kanaya S, Aoki K, Shibata D (2008) Metabolite annotations based on the integration of mass spectral information. *Plant J* 54: 949–962

Inohana Y, Yamaguchi S, Mukai N, Hirano I (2006) Development of a unique software tool to predict empirical formulae utilizing accurate mass MS^n measurements. *Chromatography* 27: 73–79 (in Japanese)

Kind T, Fiehn O (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* 7: 234

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

Nakamura Y, Kimura A, Saga H, Oikawa A, Shinbo Y, Kai K, Sakurai N, Suzuki H, Kitayama M, Shibata D, Kanaya S, Ohta D (2007) Differential metabolomics unraveling light/dark regulation of metabolic activities in *Arabidopsis* cell culture. *Planta* 227: 57–66

Oikawa A, Nakamura Y, Ogura T, Kimura A, Suzuki H, Sakurai N, Shinbo Y, Shibata D, Kanaya S, Ohta D (2006) Clarification of pathway-specific inhibition by Fourier transform ion cyclotron resonance/mass spectrometry-based metabolic phenotyping studies. *Plant Physiol* 142: 398–413

Pichersky E, Gang DR (2000) Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends Plant Sci* 5: 439–445

Rockwood AL, Haimi P (2006) Efficient calculation of accurate masses of isotopic peaks. *J Am Soc Mass Spectrom* 17: 415–419

Rockwood AL, Van Orden SL, Smith RD (1995) Rapid calculation of isotope distributions. *Anal Chem* 67: 2699–2704

Rockwood AL, Van Orman JR, Dearden DV (2004) Isotopic compositions and accurate masses of single isotopic peaks. *J Am Soc Mass Spectrom* 15: 12–21

Saito K, Hirai MY, Yonekura-Sakakibara K (2008) Decoding genes with coexpression networks and metabolomics—"majority report by precogs". *Trends Plant Sci* 13: 36–43

Senior JK (1951) Partitions and their representative graphs. *Am J Math* 73: 663–689

Stoll N, Schmidt E, Thurow K (2006) Isotope pattern evaluation for the reduction of elemental compositions assigned to high-resolution mass spectral data from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *J Am Soc Mass Spectrom* 17: 1692–1699

Suzuki H, Sasaki R, Ogata Y, Nakamura Y, Sakurai N, Kitajima M, Takayama H, Kanaya S, Aoki K, Shibata D, Saito K (2008) Metabolic profiling of flavonoids in *Lotus japonicus* using liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry. *Phytochemistry* 69: 99–111