Statistical Indices for Simultaneous Large-Scale Metabolite Detections for a Single NMR Spectrum

ARTICLE *in* ANALYTICAL CHEMISTRY · MARCH 2010

Impact Factor: 5.64 · DOI: 10.1021/ac9022023 · Source: PubMed

9 AUTHORS, INCLUDING:

Eisuke Chikayama
Niigata University of International and Info…
26 PUBLICATIONS **484** CITATIONS
SEE PROFILE

Yasuyo Sekiyama
National Food Research Institute
22 PUBLICATIONS **370** CITATIONS
SEE PROFILE

Kazuki Saito
Chiba University
494 PUBLICATIONS **18,481** CITATIONS
SEE PROFILE

Jun Kikuchi
RIKEN
154 PUBLICATIONS **3,144** CITATIONS
SEE PROFILE

Statistical Indices for Simultaneous Large-Scale Metabolite Detections for a Single NMR Spectrum

Eisuke Chikayama,[†] Yasuyo Sekiyama,[†] Mami Okamoto,[‡] Yumiko Nakanishi,[‡] Yuuri Tsuboi,[§] Kenji Akiyama,[†] Kazuki Saito,[†,||] Kazuo Shinozaki,[†] and Jun Kikuchi*[,†,‡,⊥]

Metabolome Research Group, RIKEN Plant Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, Graduate School of Nanobiosciences, Yokohama City University, 1-7-29 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, RIKEN Advanced Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan, Graduate School of Pharmaceutical Sciences, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba, Chiba 263-8522, Japan, and Graduate School of Bioagricultural Sciences, Nagoya University, 1 Furo-cho, Chikusa-ku, Nagoya-shi, Nagoya, Aichi 464-0810, Japan

NMR-based metabolomics has become a practical and analytical methodology for discovering novel genes, biomarkers, metabolic phenotypes, and dynamic cell behaviors in organisms. Recent developments in NMR-based metabolomics, however, have not concentrated on improvements of comprehensiveness in terms of simultaneous large-scale metabolite detections. To resolve this, we have devised and implemented a statistical index, the SpinAssign *p*-value, in NMR-based metabolomics for large-scale metabolite annotation and publicized this information. It enables simultaneous annotation of more than 200 candidate metabolites from the single ^{13}C-HSQC (heteronuclear single quantum coherence) NMR spectrum of a single sample of cell extract.

Metabolomics by NMR has become a promising tool that is applicable to diverse organisms, biological matters, and ecosystems. Examples are biomarkers intended for human diagnoses,[1,2] flux analyses in plants,[3,4] gut microbiomes,[5−7] genetic and transcription analyses,[8,9] and metabolic phenotyping.[10−13] Furthermore, this technique can be applied not only to living systems[14] but also to cellular-derived chemical mixtures such as food, drinks, soil samples, and water ecosystems.[15−22] Similarly, technological developments in the field of NMR-based metabolomics, recently, have become prominent (e.g., STOCSY,[23] COLMAR,[24] nonlinear sampling,[25] stable isotope-labeling techniques,[26−28] concentration quantification,[29,30] minor component

* Author to whom correspondence should be addressed. Tel.: +81-45-503-9439. Fax: +81-45-503-9489. E-mail: kikuchi@psc.riken.jp.

[†] Metabolome Research Group, RIKEN Plant Science Center.

[‡] Graduate School of Nanobiosciences, Yokohama City University.

[§] RIKEN Advanced Science Institute.

[||] Graduate School of Pharmaceutical Sciences, Chiba University.

[⊥] Graduate School of Bioagricultural Sciences, Nagoya University.

(1) Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Curr. Opin. Mol. Ther.* **2004**, *6*, 265–272.
(2) Wang, Y.; Holmes, E.; Nicholson, J. K.; Cloarec, O.; Chollet, J.; Tanner, M.; Singer, B. H.; Utzinger, J. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12676–12681.
(3) Ratcliffe, R. G.; Shachar-Hill, Y. *Plant J.* **2006**, *45*, 490–511.
(4) Sekiyama, Y.; Kikuchi, J. *Phytochemistry* **2007**, *68*, 2320–2329.
(5) Dumas, M. E.; Barton, R. H.; Toye, A.; Cloarec, O.; Blancher, C.; Rothwell, A.; Fearnside, J.; Tatoud, R.; Blanc, V.; Lindon, J. C.; Mitchell, S. C.; Holmes, E.; McCarthy, M. I.; Scott, J.; Gauguier, D.; Nicholson, J. K. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12511–12516.
(6) Wang, Y.; Cloarec, O.; Tang, H.; Lindon, J. C.; Holmes, E.; Kochhar, S.; Nicholson, J. K. *Anal. Chem.* **2008**, *80*, 1058–1066.
(7) Fukuda, S.; Nakanishi, Y.; Chikayama, E.; Ohno, H.; Hino, T.; Kikuchi, J. *PLoS One* **2009**, *4*, e4893.
(8) Hagel, J. M.; Weljie, A. M.; Vogel, H. J.; Facchini, P. *J. Plant Physiol.* **2008**, *147*, 1805–1821.
(9) Tian, C.; Chikayama, E.; Tsuboi, Y.; Kuromori, T.; Shinozaki, K.; Kikuchi, J.; Hirayama, T. *J. Biol. Chem.* **2007**, *282*, 18532–18541.
(10) Clayton, T. A.; Lindon, J. C.; Cloarec, O.; Antti, H.; Charuel, C.; Hanton, G.; Provost, J. P.; Le Net, J. L.; Baker, D.; Walley, R. J.; Everett, J. R.; Nicholson, J. K. *Nature* **2006**, *440*, 1073–1077.
(11) Slupsky, C. M.; Rankin, K. N.; Wagner, J.; Fu, H.; Chang, D.; Weljie, A. M.; Saude, E. J.; Lix, B.; Adamko, D. J.; Shah, S.; Greiner, R.; Sykes, B. D.; Marrie, T. J. *Anal. Chem.* **2007**, *79*, 6995–7004.
(12) Blaise, B. J.; Giacomotto, J.; Elena, B.; Dumas, M. E.; Toulhoat, P.; Segalat, L.; Emsley, L. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19808–19812.
(13) Smith, L. M.; Maher, A. D.; Want, E. J.; Elliott, P.; Stamler, J.; Hawkes, G. E.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 4847–4856.
(14) Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Anal. Chem.* **2003**, *75*, 384A–391A.
(15) Wishart, D. S. *Trends Food Sci. Technol.* **2008**, *19*, 482–493.
(16) Son, H. S.; Hwang, G. S.; Kim, K. M.; Kim, E. Y.; van den Berg, F.; Park, W. M.; Lee, C. H.; Hong, Y. S. *Anal. Chem.* **2009**, *81*, 1137–1145.
(17) Tiziani, S.; Schwartz, S. J.; Vodovotz, Y. *J. Agric. Food Chem.* **2006**, *54*, 6094–6100.
(18) Fan, T. W. M.; Bird, J. A.; Brodie, E. L.; Lane, A. N. *Metabolomics* **2009**, *5*, 108–122.
(19) Bundy, J. G.; Sidhu, J. K.; Rana, F.; Spurgeon, D. J.; Svendsen, C.; Wren, J. F.; Sturzenbaum, S. R.; Morgan, A. J.; Kille, P. *BMC Biol.* **2008**, *6*, 25.
(20) Viant, M. R. *Mar. Ecol.: Prog. Ser.* **2007**, *332*, 301–306.
(21) Gjersing, E. L.; Herberg, J. L.; Horn, J.; Schaldach, C. M.; Maxwell, R. S. *Anal. Chem.* **2007**, *79*, 8037–8045.
(22) Rosenblum, E. S.; Viant, M. R.; Braid, B. M.; Moore, J. D.; Friedman, C. S.; Tjeerdema, R. S. *Metabolomics* **2005**, *1*, 199–209.
(23) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77*, 1282–1289.
(24) Robinette, S. L.; Zhang, F.; Bruschweiler-Li, L.; Bruschweiler, R. *Anal. Chem.* **2008**, *80*, 3606–3611.
(25) Hyberts, S. G.; Heffron, G. J.; Tarragona, N. G.; Solanky, K.; Edmonds, K. A.; Luithardt, H.; Fejzo, J.; Chorev, M.; Aktas, H.; Colson, K.; Falchuk, K. H.; Halperin, J. A.; Wagner, G. *J. Am. Chem. Soc.* **2007**, *129*, 5108–5116.
(26) Ratcliffe, R. G.; Shachar-Hill, Y. *Biol. Rev.* **2005**, *80*, 27–43.
(27) Kikuchi, J.; Hirayama, T. *Methods Mol. Biol.* **2007**, *358*, 273–286.
(28) Ye, T.; Mo, H.; Shanaiah, N.; Gowda, G. A.; Zhang, S.; Raftery, D. *Anal. Chem.* **2009**, *81*, 4882–4888.
(29) Lewis, I. A.; Schommer, S. C.; Hodis, B.; Robb, K. A.; Tonelli, M.; Westler, W. M.; Sussman, M. R.; Markley, J. L. *Anal. Chem.* **2007**, *79*, 9385–9390.

10.1021/ac9022023 © 2010 American Chemical Society
Published on Web 02/02/2010

analysis,[31,32] new instrument applications,[33,34] and metabolite chemical shift databases[35−37]).

However, another development in terms of comprehensiveness is required for a large-scale NMR-based metabolome analysis. The above technological advancements, which are related to multidimensional NMR spectroscopy, cannot achieve large-scale chemical shift assignments, because reliably assigning large numbers of chemical shifts with multiple multidimensional NMR spectra is difficult and laborious.[38,39] Thus, current approaches cannot deal with comprehensive metabolite annotations such as simultaneous annotations for more than 100 metabolites, and such a capability is not expected for a decade.

Sophisticated methodologies and tools exist in other fields. For example, bioinformatics utilizes tools such as BLAST,[40] which is essential for classifying gene and protein sequences. These tools are widely used in genomic studies such as comparative genomics[41] and molecular phylogenetics[42] and have resulted in large, publicly available genetic and protein sequence databases[43] that have substantially contributed to the success of genomics in modern science.

Mathematical proximity indices for determining relationships between sequences, such as the E-value in BLAST, are critical to these bioinformatics approaches. To date, genomics-like indices have not been available for NMR-based metabolomics. However, some related indices are available; for example, a chemical compound can be converted to a string of characters in SMILES format.[44] Such a string can theoretically be analyzed using a genomics-like sequence alignment algorithm, and the distance between two metabolites can be computed. Another example is the structural similarity scores that can be assigned between chemical structures; using these scores, PubChem[43] can implement a clustering function for drawing a dendrogram among the given compounds. However, such indices are inappropriate for annotating a spectrum because they require structural information in advance. Although an index specifying overlaps in a spectral database has been proposed,[45] this approach does not provide statistical significance for metabolite annotation, as done by the E-value in BLAST. Hence, to improve the utility of NMR-based metabolomics, a spectrum- and mathematics-based approach for large-scale metabolite annotation is required. Such a methodology will open the door for genomics-like success in NMR-based metabolomics research.

Here, we propose such a methodology for large-scale metabolite annotation using a spectrum- and mathematics-based index called the SpinAssign *p*-value. Using this *p*-value system to analyze a single extract of ^{13}C-labeled *Arabidopsis thaliana* T87 cultured cells, we simultaneously annotated more than 200 candidate metabolites from only one ^{13}C-HSQC (heteronuclear single quantum coherence) NMR spectrum. This annotation is available on the SpinAssign server (http://prime.psc.riken.jp/?action=nmr_search).

EXPERIMENTAL SECTION

Standardized Buffer for NMR. We combined a KH_2PO_4 solution in H_2O (1 M, 1.15 mL), a K_2HPO_4 solution in H_2O (1 M, 1.85 mL), deuterium oxide (D_2O, 27.0 mL), and 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) (6.5 mg) as a chemical shift reference to make 30 mL of 100 mM potassium phosphate buffer solution (pH 7.0, 1.0 mM DSS).

Sample Preparation. We obtained T87 *Arabidopsis thaliana* cultured cells (RIKEN BioResource Center, Tsukuba, Ibaraki, Japan). Cells were incubated and subcultured every seven days in 20 mL of JPL medium,[46] supplemented with 0.5% (v/v) [$^{13}C_6$] glucose in a 100-mL baffled Erlenmeyer flask on a rotary shaker at 100 rpm and 24 °C under a 16-h-light/8-h-dark cycle. The cells were washed twice with water, lyophilized, and ground to powder. We suspended 80 mg of the powder in 600 μL of the standard NMR buffer, heated it to 65 °C for 60 min, and centrifuged it at 13 000*g* for 15 min. The supernatant was saved, and the residue was resuspended in 300 μL of the NMR buffer and treated in the same manner. We mixed and decanted both supernatants into a 5-mm-diameter NMR tube. We prepared samples for χ^2 distribution validations and the reference chemical shift database, as described elsewhere.[47]

NMR Experiments and Analyses. We acquired a ^{13}C-HSQC spectrum of ^{13}C-labeled T87 cultured cell extract with a

[30] Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M. *Anal. Chem.* **2006**, *78*, 4430–4442.

[31] Sandusky, P.; Raftery, D. *Anal. Chem.* **2005**, *77*, 2455–2463.

[32] Weljie, A. M.; Newton, J.; Jirik, F. R.; Vogel, H. J. *Anal. Chem.* **2008**, *80*, 8956–8965.

[33] Keun, H. C.; Beckonert, O.; Griffin, J. L.; Richter, C.; Moskau, D.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2002**, *74*, 4588–4593.

[34] Krojanski, H. G.; Lambert, J.; Gerikalan, Y.; Suter, D.; Hergenroder, R. *Anal. Chem.* **2008**, *80*, 8668–8672.

[35] Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Wenger, R. K.; Yao, H. Y.; Markley, J. L. *Nucleic Acids Res.* **2008**, *36*, D402–D408.

[36] Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35*, D521–526.

[37] Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L. *Nat. Biotechnol.* **2008**, *26*, 162–164.

[38] Nicholson, J. K.; Foxall, P. J.; Spraul, M.; Farrant, R. D.; Lindon, J. C. *Anal. Chem.* **1995**, *67*, 793–811.

[39] Fan, W. M. T. *Prog. Nucl. Magn. Reson. Spectrosc.* **1996**, *28*, 161–219.

[40] Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

[41] Carlton, J. M.; Adams, J. H.; Silva, J. C.; Bidwell, S. L.; Lorenzi, H.; Caler, E.; Crabtree, J.; Angiuoli, S. V.; Merino, E. F.; Amedeo, P.; Cheng, Q.; Coulson, R. M. R.; Crabb, B. S.; del Portillo, H. A.; Essien, K.; Feldblyum, T. V.; Fernandez-Becerra, C.; Gilson, P. R.; Gueye, A. H.; Guo, X.; Kang'a, S.; Kooij, T. W. A.; Korsinczky, M.; Meyer, E. V. S.; Nene, V.; Paulsen, I.; White, O.; Ralph, S. A.; Ren, Q. H.; Sargeant, T. J.; Salzberg, S. L.; Stoeckert, C. J.; Sullivan, S. A.; Yamamoto, M. M.; Hoffman, S. L.; Wortman, J. R.; Gardner, M. J.; Galinski, M. R.; Barnwell, J. W.; Fraser-Liggett, C. M. *Nature* **2008**, *455*, 757–763.

[42] Murphy, W. J.; Eizirik, E.; Johnson, W. E.; Zhang, Y. P.; Ryderk, O. A.; O'Brien, S. J. *Nature* **2001**, *409*, 614–618.

[43] Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; DiCuccio, M.; Edgar, R.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Miller, V.; Ostell, J.; Pruitt, K. D.; Schuler, G. D.; Shumway, M.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. *Nucleic Acids Res.* **2008**, *36*, D13–D21.

[44] Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[45] Xia, J. G.; Bjorndahl, T. C.; Tang, P.; Wishart, D. S. *BMC Bioinf.* **2008**, *9*, 507.

[46] Axelos, M.; Curie, C.; Mazzolini, L.; Bardet, C.; Lescure, B. *Plant Physiol. Biochem.* **1992**, *30*, 123–128.

[47] Chikayama, E.; Suto, M.; Nishihara, T.; Shinozaki, K.; Kikuchi, J. *PLoS ONE* **2008**, *3*, e3805.

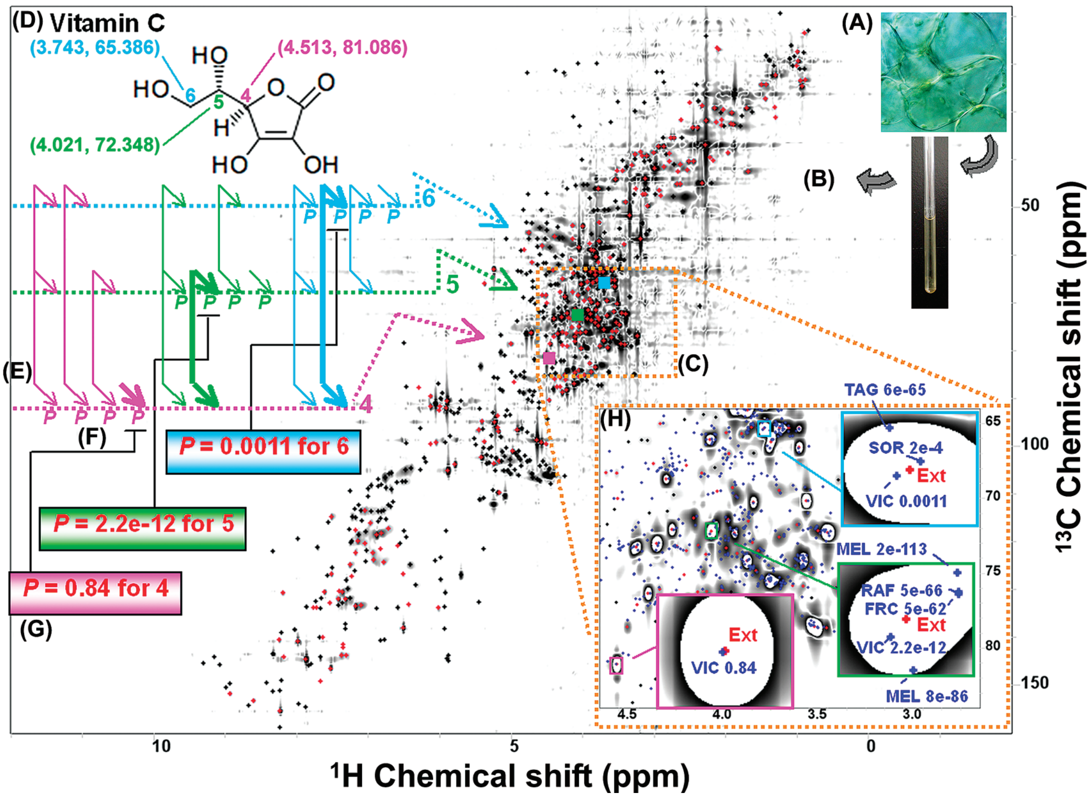

Figure 1. SpinAssign *p*-value methodology: (A) ^{13}C-labeled *Arabidopsis thaliana* T87 cultured cells for extraction; (B) ^{13}C-HSQC spectrum with 649 peaks (red and black crosses), of which 272 (red crosses) are candidates in our database; (C, D) three annotations for Vitamin C (carbon 4 (pink), 5 (green), and 6 (light blue)); (E) four combined groups of solid arrows (pink); (F) four standard *p*-values (*P* values given in pink); (G) *p*-value of 0.84 for annotation 4 (thick arrow shown in pink); (H) an expanded region (dotted in orange). (Reference database peaks with *p*-values (Vitamin C (VIC), melibiose (MEL), D-fructose (FRC), raffinose (RAF), D-sorbitol (SOR), and tagatose (TAG)). Database-matching areas (0.06 ppm × 1.06 ppm boxes colored pink, green, and light blue) for the user peaks (denoted as “Ext”).

resolution of 2048 points per 20 ppm in the ^{1}H dimension, accumulating 160 transients per free induction decay (FID) and 720 increments per 220 ppm in the ^{13}C dimension at 298 K on an NMR spectrometer (Bruker AVANCE 700) equipped with a ^{1}H inverse cryogenic probe with triple-axis gradients, operating at 700.15 MHz. The spectrum was processed with the NMRPipe software package.[48] We used exponential windows with line-broadening parameters of 10 and 15 Hz in the ^{1}H and ^{13}C dimensions, respectively, 54° phase-shifted sinbell windows, zero fillings, and polynomial baseline corrections in both dimensions. The automated peak-picking function in the NMRDraw software (NMRPipe package) picked 2370 peaks initially after eliminating negative peaks. This number was reduced to 1523 peaks with a customized program that roughly filters decoupling sideband artifacts and baseline artifacts in a ^{13}C-HSQC spectrum (available at the SpinAssign website). Finally, 649 curated peaks were manually obtained by carefully deleting and appending peaks. We submitted these peaks to SpinAssign with tolerance parameters of 0.03 and 0.53 ppm in the ^{1}H and ^{13}C dimensions, respectively, and obtained annotations with *p*-values. NMR experiments for χ^2 distribution validations and the reference chemical shift database were performed as described elsewhere.[47]

Functionality and System Design. The functionality of this annotation is available at http://prime.psc.riken.jp/?action=nmr_search. Its function is large-scale metabolite annotation. It accepts a user query that consists of ^{1}H and ^{13}C chemical shifts (^{13}C-HSQC peaks) in text format. Queried peaks are matched against the reference chemical shift database with tolerances for ^{1}H and ^{13}C. Although a user can specify tolerances, default values are empirically tuned to ±0.03 and ±0.53 ppm for the ^{1}H and ^{13}C dimensions, respectively. After the queried peaks are matched against the database and the result is transferred back to the user’s client browser, all *p*-values are computed and displayed on the client side. An arbitrary precision library in JavaScript was used to develop the *p*-value computing program. The *p*-values on the website are colored red (p-value $= 1$) to black (p-value $= 1 \times 10^{-62}$); the latter is the point at which two calculations with different tolerances do not yield the same results. The SpinAssign system was developed with HTML; JavaScript was used for the *p*-value computation program; PHP, a server-side programming language, was used for server-side programs; and the database was managed with MySQL, a relational database system, which implemented the reference chemical shift database including more than 1700 ^{13}C-HSQC peaks, corresponding to 270 metabolites.

Uniqueness. Uniqueness is the extent to which a peak in the reference chemical shift database does not overlap with other nearby peaks in the reference chemical shift database. Uniqueness for a reference database peak is defined as

$$\text{uniqueness} = \frac{1}{C}$$

[48] Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277–293.

where C is the number of matches around the reference peak when it is queried to the database using the tolerance parameters of 0.03 and 0.53 ppm for the ^{1}H and ^{13}C dimensions, respectively. The values range from 1 to 0, where 1 means no overlap and 0 means infinite overlaps. The queried peak itself is always matched and, therefore, C never exceeds 1.

RESULTS AND DISCUSSION

SpinAssign p-values represent proximities between user and reference HSQC peaks of metabolites and are computed from the gaps between them. These gaps are assumed to follow Gaussian distribution and, therefore, can be deduced by the χ^2 distribution for the total sum of the squared gaps involved. The standard p-value for the χ^2 distribution in statistics is the cumulative χ^2 distribution subtracted from 1. SpinAssign p-values are calculated similarly, and they target not only a single standard p-value but also combined multiple standard p-values (for the mathematical definition, see the Supporting Information). SpinAssign p-values fall in the range of 1 to 0, where 1 indicates no gaps (peaks completely match) and 0 indicates that one or more gaps are infinitely large (peaks are completely different). The notations of SpinAssign and standard p-values are not differentiated in the following text, unless necessary for clarity.

The following steps demonstrate the SpinAssign p-value calculation for Vitamin C, which is an essential metabolite in both humans and plants (see Figure 1):

(1) We prepared an NMR tube of extract (Figure 1A).

(2) A ^{13}C-HSQC spectrum was recorded, and peaks were picked, curated, and annotated with our in-house database-matching program (Figure 1B).

(3) We annotated the three found user peaks (Figure 1C) as atoms in Vitamin C (Figure 1D).

(4) All the possible combinations of one or more annotations including annotation 4, formed as four groups, were computed (Figure 1E).

(5) Standard p-values, interpreted as the probability of the coincident occurrence of the combinations against each of the four groups, were calculated (Figure 1F) using the χ^2 distribution with the number of degrees of freedom equal to the number of combined peaks within a group, multiplied by 2, obtained from the dimensions of the spectra.

(6) Finally, we assigned the SpinAssign p-value for this annotation as the maximum of the four standard p-values (Figure 1G). Similarly, we computed the p-values for annotations 5 and 6. Although the p-values of candidate metabolites other than Vitamin C were found for the three user peaks, the highest observed p-values were for Vitamin C (the larger the gaps, the smaller the p-values; Figure 1H). However, it is essential that other potential metabolites be statistically quantified simultaneously via this method.

To confirm our algorithm, we examined whether metabolites in a common standard condition exhibit the χ^2 distribution (see Figure 2). We statistically analyzed standard compounds and extracts of cultured cells in solvent ($N = 7$) and determined the standard deviation parameters involved in the computations of p-values to be 0.00177 and 0.0244 ppm for ^{1}H and ^{13}C, respectively. The χ^2 density distribution fluctuates because the number of combinations of standard compounds is small. Nonetheless, the χ^2 cumulative distribution, which is used for p-value calculation, agreed well with the theoretical curves.

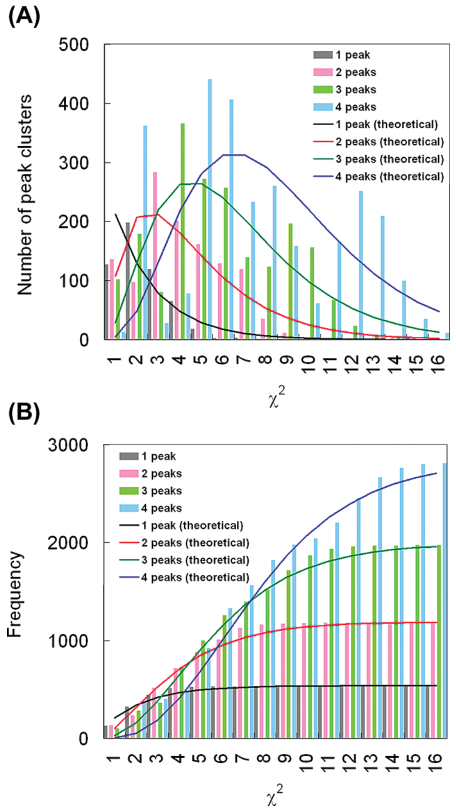

Figure 2. χ^2 density and cumulative distributions of 19 standard compounds in a standard buffer to which the cultured cell extract was added: (A) observed χ^2 density distributions (one (gray, $m = 1$), two (pink, $m = 2$), three (green, $m = 3$), and four (light blue, $m = 4$) gaps per standard compound were selected for all combinations and counted; theoretical χ^2 density distributions (solid lines) are also shown); and (B) corresponding cumulative distributions. The standard compounds (16 protein amino acids,excluding asparagine, glutamine, cystein, and tryptophan; glucose; succinate; and fumarate) were dissolved seven times in a buffer added with a different lot of the extract. For each observed ^{13}C-HSQC peak of the standard compounds, gaps from the corresponding average over the seven experiments were determined. We expected that the gaps would be Gaussian-distributed and the sum of m different squared gaps would generate the χ^2 density distribution with the number of degrees of freedom equal to $m \times 2$ (because a ^{13}C-HSQC peak has two degrees of freedom). Our SpinAssign p-value definition demands that the m-times selection be performed for each standard compound. The pH for all samples in all the seven experiments exhibited slight fluctuations (7.2 ± 0.1).

Applying our methodology to the NMR spectrum of a ^{13}C-labeled *Arabidopsis* T87 cultured-cell extract (see Figures 1A and 1B), we simultaneously annotated more than 200 metabolites. Figure 3 shows the p-value, uniqueness, and ratio of the number of detected peaks to that of all peaks for each metabolite (detection rate). Uniqueness is the extent of overlap with a reference peak in the reference database (see Figure S1 in the Supporting Information). Its values fall in the range of 1−0, where 1 indicates no overlap and 0 indicates infinite overlaps. Multiple high p-values are most reliable. While a 100% detection rate indicates high reliability in nonlabeled samples, a lower detection rate in a ^{13}C-labeled sample may still be acceptable, because the detection rate can be less than 100%, despite intense metabolite peaks if

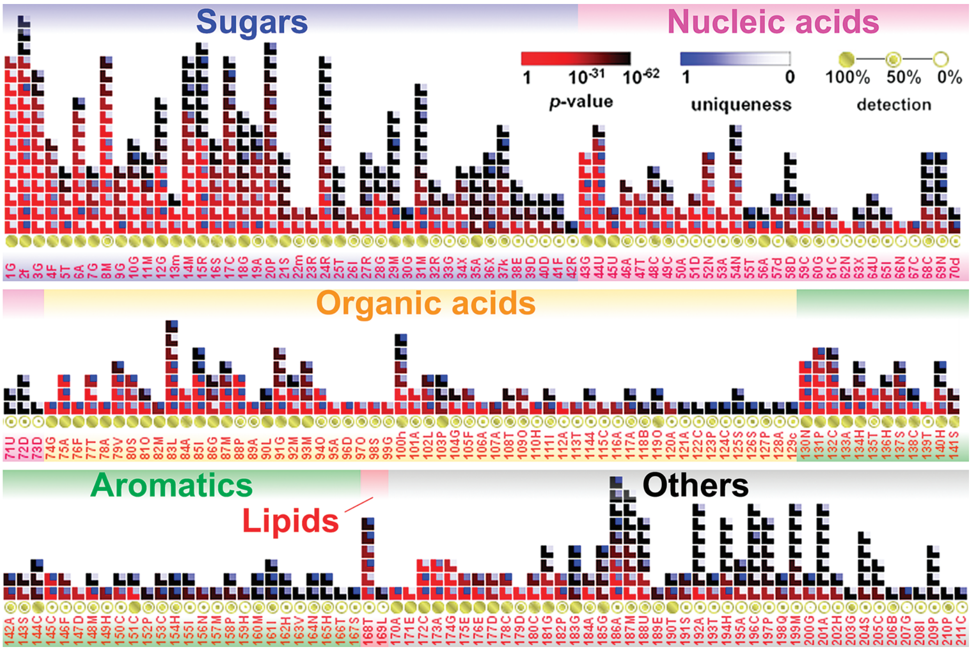

Figure 3. All detected annotations summing to 211 metabolites in the NMR spectrum of *Arabidopsis* T87 cultured cell extract. Annotations per metabolite are stacked vertically. For a given metabolite, each block in a column represents one annotation (presumed peak) for that metabolite. Three values are shown for each metabolite: the *p*-value for each peak (the block color ranges from red to black or the *p*-value = 1 to 1 × 10^{-62}), uniqueness (small inset block ranges from blue to white), and the detection rate (the size of the green inner circle at the base of each metabolite column). Sequential numbers and a representative letter are aligned horizontally for all candidate metabolites. (See Table S1 in the Supporting Information).

carbon atoms for the metabolite are partly labeled. In total, 42 sugars, 31 nucleic acids, 56 organic acids, 38 aromatics, 2 lipids, and 42 other metabolites were annotated with our methodology, demonstrating the advantage of NMR-based large-scale metabolite annotation in terms of chemical comprehensiveness and uniform profiling. However, we should note that a mass-spectrometry-based approach is chemically more selective.[49] Generally, such a metabolic profile is affected by experimental protocol, including the type of extraction solvent. Extraction methods are important in metabolomics. Protocol optimization (e.g., the use of an organic solvent for recovering lipids or cold perchloric acid for quenching the progress of enzymatic reactions[50,51]) is encouraged. In our result, because of the use of an aqueous buffer, annotations for lipids are limited, because they are not easily recovered.

Our method has been made available on the improved SpinAssign server (http://prime.psc.riken.jp/?action=nmr_search) on the PRIMe website.[52] SpinAssign annotation results are linked to KEGG.[53] A standard protocol and procedure for data collection and for using the tool to assign *p*-values are available in the Supporting Information.

We caution that SpinAssign annotation does not provide absolute proof of the existence of annotated metabolites. This is mainly due to overlaps of reference database peaks in chemical shifts (which affect up to 70% of all reference peaks; see Figure S1B in the Supporting Information). Although our *p*-value method targets the overlaps, *p*-value thresholds for reliable annotation should be empirically determined for each project on the basis of a series of experiments of the same type. This is because chemical shifts of metabolites dissolved in an extract are perturbed from those in the reference buffer, because of metabolites or ions extracted from sample cells. For example, in our result, *p*-values above 1 × 10^{-20} are regarded as highly reliable; those below 1 × 10^{-20} and above 1 × 10^{-50} are considered to be in a gray zone (not apparent); and those below 1 × 10^{-50} are regarded as barely reliable (see the Supporting Information for further discussion). We expect similar thresholds for similar experiments. Note that we claim that all the annotations in our result (Figure 3) are only screening candidates; hence, users who expect greater reliability from NMR, as compared to SpinAssign, should acquire and analyze additional multidimensional NMR spectra.[39,47] Concomitant use of other annotation websites[35,36] may also be effective, although no other system uses the *p*-values reported here.

Although our new methodology and tools result in a mix of accurate and inaccurate information, because of the comprehensive use of *p*-values, ambiguous information as in genomics E-values can still be biologically interpreted. First, it is theoretically correct that a higher *p*-value demonstrates a higher probability of occurrence of a common chemical substructure. In metabolic reactions and pathways, metabolites are divided, converted, and combined step-by-step. These pathways lead to common substructures before and after reactions. As a result, our methodology is necessarily applicable to metabolic studies. Second, a heterocorrelation between a metabolome and a genome, transcriptome, or

(49) Pan, Z.; Raftery, D. *Anal Bioanal. Chem.* **2007**, *387*, 525–527.

(50) Kruger, N. J.; Troncoso-Ponce, M. A.; Ratcliffe, R. G. *Nat. Protocols* **2008**, *3*, 1001–1012.

(51) Fan, T. W. M.; Lane, A. N. *Prog. Nucl. Magn. Reson. Spectrosc.* **2008**, *52*, 69–117.

(52) Akiyama, K.; Chikayama, E.; Yuasa, H.; Shimada, Y.; Tohge, T.; Shinozaki, K.; Hirai, M. Y.; Sakurai, T.; Kikuchi, J.; Saito, K. *In Silico Biol.* **2008**, *8*, 339–345.

(53) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. *Nucleic Acids Res.* **2006**, *34*, D354–D357.

proteome analysis[54,55] can screen functional genes or proteins, using annotated candidate metabolite names, if they acquire common substructures along their pathways. Third, a holistic metabolic pattern derived from our *p*-value methodology essentially represents a new type of data regarding the metabolic properties of a sample. In general, the behavior of each metabolite and the holistic relationships among metabolite behaviors in an organism are physically independent variables and are determined independently. One example is the study of the coarse-grained view of metabolic pathways in silkworms, demonstrating a visualized global metabolic pattern (relationships among metabolites and pathways).[47] These facts support the need for a holistic methodology (although inaccurate information is mixed).

[54] Ishii, N.; Nakahigashi, K.; Baba, T.; Robert, M.; Soga, T.; Kanai, A.; Hirasawa, T.; Naba, M.; Hirai, K.; Hoque, A.; Ho, P. Y.; Kakazu, Y.; Sugawara, K.; Igarashi, S.; Harada, S.; Masuda, T.; Sugiyama, N.; Togashi, T.; Hasegawa, M.; Takai, Y.; Yugi, K.; Arakawa, K.; Iwata, N.; Toya, Y.; Nakayama, Y.; Nishioka, T.; Shimizu, K.; Mori, H.; Tomita, M. *Science* **2007**, *316*, 593–597.

[55] Mochida, K.; Furuta, T.; Ebana, K.; Shinozaki, K.; Kikuchi, J. *BMC Genomics* **2009**, *10*, 568.

ACKNOWLEDGMENT

We thank H. Haruna (Yokohama City University) and Y. Tosaka (RIKEN PSC) for improving the chemical shift database and R. C. Everroad (RIKEN) for manuscript editing. This work was partly supported by grants from CREST, the Japan Science and Technology Agency (J.K.), the Research and Development Program for New Bioindustry Initiatives of the Bio-oriented Technology Research Advancement Institution (BRAIN to J.K.), and the New Energy and Industrial Technology Development Organization (NEDO to J.K.).

SUPPORTING INFORMATION AVAILABLE

Additional information is available as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review September 30, 2009. Accepted January 4, 2010.

AC9022023