PO-CON1307E

Automated compound identification using product ion scanning with accurate mass measurement and compound database searching for non-targeted metabolomics

ASMS 2013 MP03-044

Tairo Ogura[1,2]; Faith Hays[3]; Takeshi Bamba[1];
Eiichiro Fukusaki[1]
[1]Graduate School of Engineering, Osaka University,
Osaka, JAPAN;
[2]Shimadzu corporation, Kyoto , JAPAN;
[3]Shimadzu Scientific Instruments, Columbia, MD

1. Introduction

Non-targeted metabolomics entails data exploration to find the important metabolites from the detected features. Liquid chromatography mass spectrometry (LC-MS) is frequently used for non-targeted metabolomics because of the analytical sensitivity for a wide variety of compounds. In non-targeted metabolomics, compound identification is an important step to translate the information obtained from the instrument such as retention time and *m/z* into biologically relevant information such as chemical name and structure. We therefore developed a compound identification technique with scoring using both prediction formulae and assignment of product ions for narrowing the candidates. Using this approach, each candidate is evaluated not only using partial structural information from the spectral assignment, but also using all the molecular information provided by formula prediction.

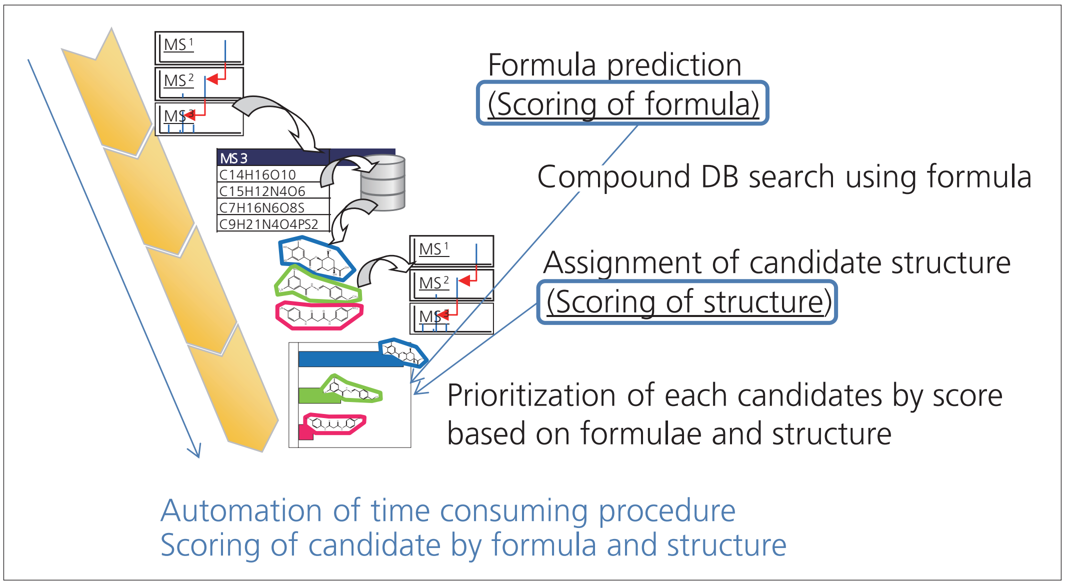

MS1
MS2
MS3
MS 3
C14H16O10
C15H12N4O6
C7H16N6O8S
C9H21N4O4PS2
MS1
MS2
MS3
Formula prediction
(Scoring of formula)
Compound DB search using formula
Assignment of candidate structure
(Scoring of structure)
Prioritization of each candidates by score
based on formulae and structure
Automation of time consuming procedure
Scoring of candidate by formula and structure

Fig. 1 Schematic view of compound identification system

2. Methods and Materials

To evaluate the developed technique, we used dried green tea leaves which had been ranked by a sensory evaluation test. An aliquot of extracted samples was injected into an LCMS-IT-TOF system (Shimadzu Co.) with an ESI source. We applied this technique to find compounds which are important to the quality of the tea by constructing a quality prediction model using multivariate analysis. Formula Predictor (Shimadzu Co.) was used for formula prediction. These formulae were then used for database searching by an in-house developed searching interface and Application Programming Interface derived from ChemSpider. After predicting the list of candidate compounds, the score for each candidate was calculated based on mass accuracy and comparison of observed and predicted tandem mass spectra.

Table 1 Analytical conditions

column	: Shim-pack XR-ODS (2.0 mm I.D. × 50 mm L., 2.2 µm)		
mobile phase A	: Water containing 0.1% formic acid		
mobile phase B	: Methanol		
gradient program	: 2%B (0 min) – 60%B (10 min) – 98%B (10.01-14 min) – 2%B (14.01 – 19 min)		
flow rate	: 0.4 mL/min	column temp.	: 40°C
ionization	: ESI (+/- switching)	scan range	: *m/z* 100 – 1000
CDL temp.	: 200°C	BH temp.	: 200°C

3. Result

3-1. Chromatograms of green tea extract

3742 peaks were detected from tea extract. These peaks were narrowed to 462 by filtering of isotopic peaks and p-value. This peak set was used to construction of tea quality evaluation model.

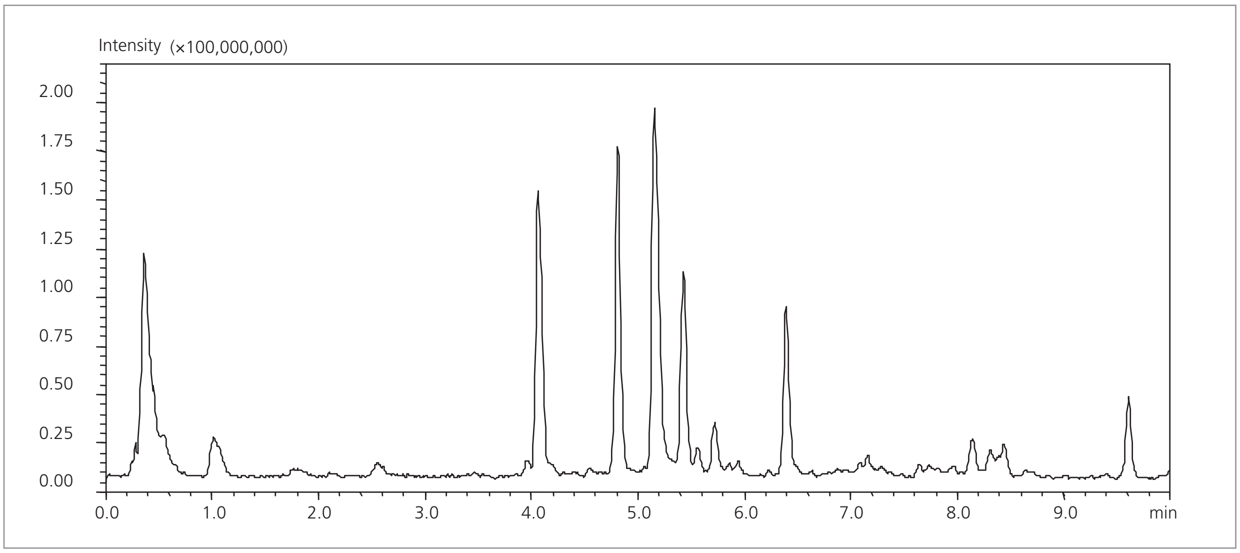


Fig. 2 Typical TIC chromatogram of green tea leaf extract.

3-2. Construction of tea quality evaluation model

To select compounds which shows importance to tea quality, quality evaluation model using PLS regression model were constructed. Compound identification for top 20 compounds which shows highest impact in variable importance in the projection plot were performed.

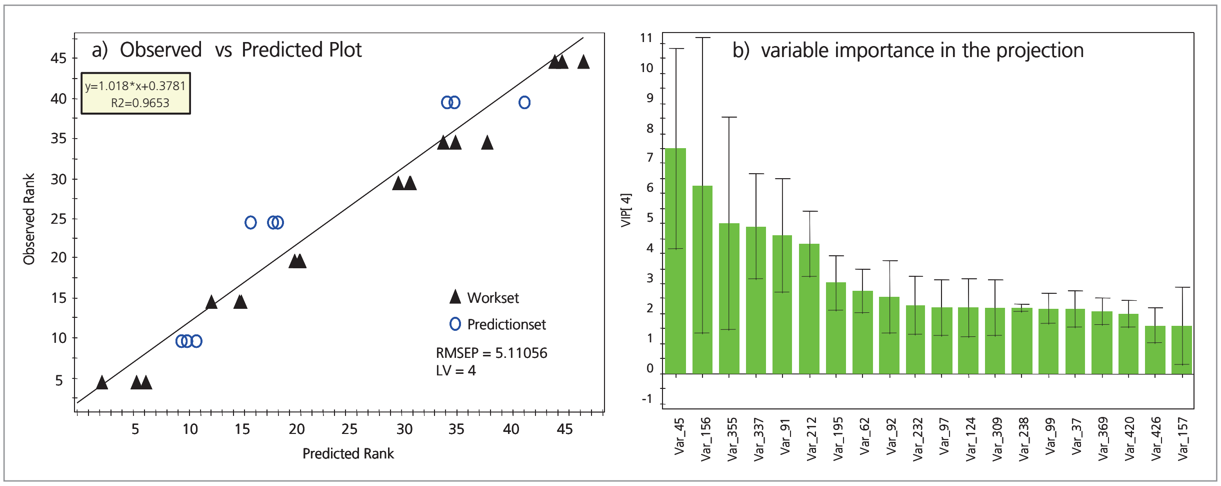
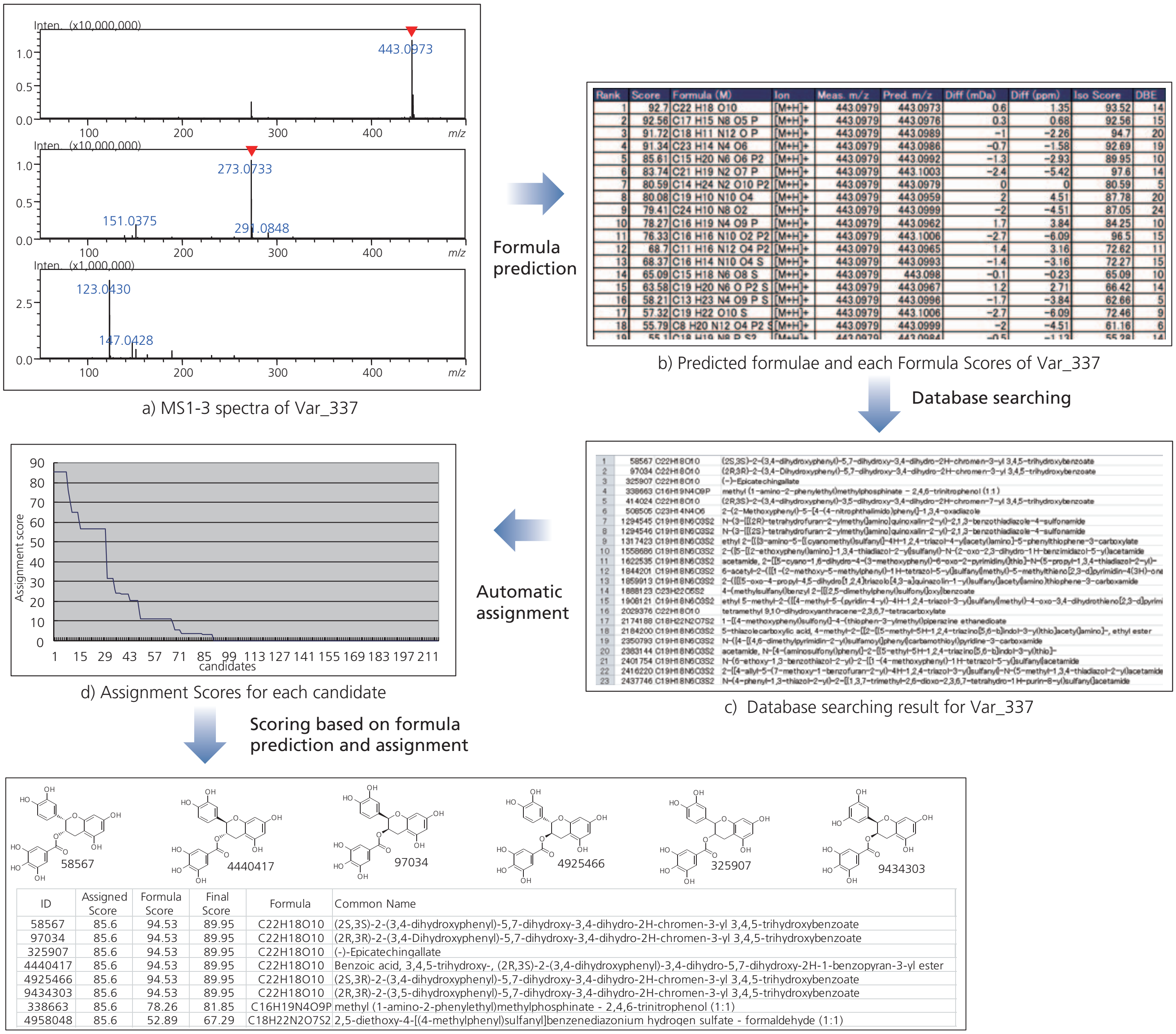


Fig. 3 Quality evaluation model of green tea

3-3. Compound identification

The result of compound identification for var_337 was shown here as an example. Formula prediction for var_337 was performed using MS$^{1-3}$ spectra (Figs. 4a-b). The score of the chemical formula (Formula Score) was calculated based on comparison of theoretical and observed *m/z* value and isotopic patterns using Formula Predictor. By using this formula list as a query for database searching, 218 candidate compounds were retrieved (Fig. 4c). The score of the assignment (Assignment Score) was calculated based on rate of assigned ion among product ion spectrum. As a result of automatic assignment of product ion spectrum, eight candidates received a highest score (Fig. 4d). Finally, 218 candidates were narrowed to 6 candidates by the scoring based on formula prediction and automatic assignment (Fig. 4e).



a) MS1-3 spectra of Var_337

Rank	Score	Formula (M)	Ion	Meas. m/z	Pred. m/z	Diff (mDa)	Diff (ppm)	Iso Score	DBE
1	92.7	C22 H18 O10	[M+H]+	443.0979	443.0973	0.6	1.35	93.52	14
2	92.56	C17 H15 N8 O5 P	[M+H]+	443.0979	443.0976	0.3	0.68	92.56	15
3	91.72	C18 H11 N12 O P	[M+H]+	443.0979	443.0989	−1	−2.26	94.7	20
4	91.34	C23 H14 N4 O6	[M+H]+	443.0979	443.0986	−0.7	−1.58	92.69	19
5	85.61	C15 H20 N6 O6 P2	[M+H]+	443.0979	443.0992	−1.3	−2.93	89.95	10
6	83.74	C21 H19 N2 O7 P	[M+H]+	443.0979	443.1003	−2.4	−5.42	97.6	14
7	80.59	C14 H24 N2 O10 P2	[M+H]+	443.0979	443.0979	0	0	80.59	5
8	80.08	C19 H10 N10 O4	[M+H]+	443.0979	443.0959	2	4.51	87.78	20
9	79.41	C24 H10 N8 O2	[M+H]+	443.0979	443.0999	−2	−4.51	87.05	24
10	78.27	C16 H19 N4 O9 P	[M+H]+	443.0979	443.0962	1.7	3.84	84.25	10
11	76.33	C16 H16 N10 O2 P2	[M+H]+	443.0979	443.1006	−2.7	−6.09	96.5	15
12	68.7	C11 H16 N12 O4 P2	[M+H]+	443.0979	443.0965	1.4	3.16	72.62	11
13	68.37	C16 H14 N10 O4 S	[M+H]+	443.0979	443.0993	−1.4	−3.16	72.27	15
14	65.09	C15 H18 N6 O8 S	[M+H]+	443.0979	443.098	−0.1	−0.23	65.09	10
15	63.58	C19 H20 N6 O P2 S	[M+H]+	443.0979	443.0967	1.2	2.71	66.42	14
16	58.21	C13 H23 N4 O9 P S	[M+H]+	443.0979	443.0996	−1.7	−3.84	62.66	5
17	57.32	C19 H22 O10 S	[M+H]+	443.0979	443.1006	−2.7	−6.09	72.46	9
18	55.79	C8 H20 N12 O4 P2 S	[M+H]+	443.0979	443.0999	−2	−4.51	61.16	6
19	55.1	C18 H19 N8 P S2	[M+H]+	443.0979	443.0984	−0.5	−1.13	55.28	14

b) Predicted formulae and each Formula Scores of Var_337

1	58567	C22H18O10	(2S,3S)-2-(3,4-dihydroxyphenyl)-5,7-dihydroxy-3,4-dihydro-2H-chromen-3-yl 3,4,5-trihydroxybenzoate
2	97034	C22H18O10	(2R,3R)-2-(3,4-Dihydroxyphenyl)-5,7-dihydroxy-3,4-dihydro-2H-chromen-3-yl 3,4,5-trihydroxybenzoate
3	325907	C22H18O10	(−)-Epicatechingallate
4	338663	C16H19N4O9P	methyl (1-amino-2-phenylethyl)methylphosphinate - 2,4,6-trinitrophenol (1:1)
5	414024	C22H18O10	(2R,3S)-2-(3,4-dihydroxyphenyl)-3,5-dihydroxy-3,4-dihydro-2H-chromen-7-yl 3,4,5-trihydroxybenzoate
6	508505	C23H14N4O6	2-(2-Methoxyphenyl)-5-[4-(4-nitrophthalimido)phenyl]-1,3,4-oxadiazole
7	1294545	C19H18N6O3S2	N-(3-{[(2R)-tetrahydrofuran-2-ylmethyl]amino}quinoxalin-2-yl)-2,1,3-benzothiadiazole-4-sulfonamide
8	1294546	C19H18N6O3S2	N-(3-{[(2S)-tetrahydrofuran-2-ylmethyl]amino}quinoxalin-2-yl)-2,1,3-benzothiadiazole-4-sulfonamide
9	1317423	C19H18N6O3S2	ethyl 2-[({3-amino-5-[(cyanomethyl)sulfanyl]-4H-1,2,4-triazol-4-yl}acetyl)amino]-5-phenylthiophene-3-carboxylate
10	1558686	C19H18N6O3S2	2-({5-[(2-ethoxyphenyl)amino]-1,3,4-thiadiazol-2-yl}sulfanyl)-N-(2-oxo-2,3-dihydro-1H-benzimidazol-5-yl)acetamide
11	1622535	C19H18N6O3S2	acetamide, 2-[[5-cyano-1,6-dihydro-4-(3-methoxyphenyl)-6-oxo-2-pyrimidinyl]thio]-N-(5-propyl-1,3,4-thiadiazol-2-yl)-
12	1844201	C19H18N6O3S2	6-acetyl-2-({[1-(2-methoxy-5-methylphenyl)-1H-tetrazol-5-yl]sulfanyl}methyl)-5-methylthieno[2,3-d]pyrimidin-4(3H)-one
13	1859913	C19H18N6O3S2	2-({[(5-oxo-4-propyl-4,5-dihydro[1,2,4]triazolo[4,3-a]quinazolin-1-yl)sulfanyl]acetyl}amino)thiophene-3-carboxamide
14	1888123	C23H22O5S2	4-(methylsulfanyl)benzyl 2-{[(2,5-dimethylphenyl)sulfonyl]oxy}benzoate
15	1908121	C19H18N6O3S2	ethyl 5-methyl-2-({[4-methyl-5-(pyridin-4-yl)-4H-1,2,4-triazol-3-yl]sulfanyl}methyl)-4-oxo-3,4-dihydrothieno[2,3-d]pyrimi
16	2029376	C22H18O10	tetramethyl 9,10-dihydroxyanthracene-2,3,6,7-tetracarboxylate
17	2174188	C18H22N2O7S2	1-[(4-methoxyphenyl)sulfonyl]-4-(thiophen-3-ylmethyl)piperazine ethanedioate
18	2184200	C19H18N6O3S2	5-thiazolecarboxylic acid, 4-methyl-2-[[2-[(5-methyl-5H-1,2,4-triazino[5,6-b]indol-3-yl)thio]acetyl]amino]-, ethyl ester
19	2350793	C19H18N6O3S2	N-({4-[(4,6-dimethylpyrimidin-2-yl)sulfamoyl]phenyl}carbamothioyl)pyridine-3-carboxamide
20	2383144	C19H18N6O3S2	acetamide, N-[4-(aminosulfonyl)phenyl]-2-[(5-ethyl-5H-1,2,4-triazino[5,6-b]indol-3-yl)thio]-
21	2401754	C19H18N6O3S2	N-(6-ethoxy-1,3-benzothiazol-2-yl)-2-{[1-(4-methoxyphenyl)-1H-tetrazol-5-yl]sulfanyl}acetamide
22	2416220	C19H18N6O3S2	2-{[4-allyl-5-(7-methoxy-1-benzofuran-2-yl)-4H-1,2,4-triazol-3-yl]sulfanyl}-N-(5-methyl-1,3,4-thiadiazol-2-yl)acetamide
23	2437746	C19H18N6O3S2	N-(4-phenyl-1,3-thiazol-2-yl)-2-[(1,3,7-trimethyl-2,6-dioxo-2,3,6,7-tetrahydro-1H-purin-8-yl)sulfanyl]acetamide

c) Database searching result for Var_337

d) Assignment Scores for each candidate

ID	Assigned Score	Formula Score	Final Score	Formula	Common Name
58567	85.6	94.53	89.95	C22H18O10	(2S,3S)-2-(3,4-dihydroxyphenyl)-5,7-dihydroxy-3,4-dihydro-2H-chromen-3-yl 3,4,5-trihydroxybenzoate
97034	85.6	94.53	89.95	C22H18O10	(2R,3R)-2-(3,4-Dihydroxyphenyl)-5,7-dihydroxy-3,4-dihydro-2H-chromen-3-yl 3,4,5-trihydroxybenzoate
325907	85.6	94.53	89.95	C22H18O10	(-)-Epicatechingallate
4440417	85.6	94.53	89.95	C22H18O10	Benzoic acid, 3,4,5-trihydroxy-, (2R,3S)-2-(3,4-dihydroxyphenyl)-3,4-dihydro-5,7-dihydroxy-2H-1-benzopyran-3-yl ester
4925466	85.6	94.53	89.95	C22H18O10	(2S,3R)-2-(3,4-dihydroxyphenyl)-5,7-dihydroxy-3,4-dihydro-2H-chromen-3-yl 3,4,5-trihydroxybenzoate
9434303	85.6	94.53	89.95	C22H18O10	(2R,3R)-2-(3,5-dihydroxyphenyl)-5,7-dihydroxy-3,4-dihydro-2H-chromen-3-yl 3,4,5-trihydroxybenzoate
338663	85.6	78.26	81.85	C16H19N4O9P	methyl (1-amino-2-phenylethyl)methylphosphinate - 2,4,6-trinitrophenol (1:1)
4958048	85.6	52.89	67.29	C18H22N2O7S2	2,5-diethoxy-4-[(4-methylphenyl)sulfanyl]benzenediazonium hydrogen sulfate - formaldehyde (1:1)

Fig. 4 Compound identification result for Var_337

Table 2 Candidates representing the 20 components that most impact tea quality as determined by VIP value

ID	*m/z*	R.T. (min)	Candidate	Formula	Ion	Final Score
UK -001	195.087	4.807	caffeine	$C_8H_{10}N_4O_2$	[M+H]+	94.24
UK -002	307.080	4.063	gallocatechin	$C_{15}H_{14}O_7$	[M+H]+	87.86
UK -003	459.092	5.156	gallocatechin gallate	$C_{22}H_{18}O_{11}$	[M+H]+	93.07
UK -004	443.097	6.392	catechin gallate	$C_{22}H_{18}O_{10}$	[M+H]+	89.95
UK -005	261.169	5.429	1-(4-amino-6,7,8,9-tetrahydro -1h-imidazo[4,5-c]quinolin -1-yl) -2-methylpropan -2-ol	$C_{14}H_{20}N_4O$	[M+H]+	73.06
UK -006	345.080	1.003	theogalline	$C_{14}H_{16}O_{10}$	[M+H]+	93.84
UK -007	339.106	5.719	coumaroyl quinic acid	$C_{16}H_{18}O_8$	[M+H]+	82.14
UK -008	217.068	4.805	(Sodium ion adduct of UK-001)			
UK -009	261.169	4.034	1-(4-amino-6,7,8,9 -tetrahydro -1h-imidazo[4,5-c]quinolin -1-yl) -2-methylpropan -2-ol	$C_{14}H_{20}N_4O$	[M+H]+	75.44
UK -010	361.088	5.717	(Sodium ion adduct of UK-007)			
UK -011	273.074	6.393	(Fragment of UK -004)			
UK -012	365.159	4.4	ethyl -5-(acetylamino) -2,3,4,5-tetradeoxy-2-methylidene -4-nitro-d-glycero-d-galacto-nononate	$C_{14}H_{24}N_2O_9$	[M+H]+	58.90
UK -013	291.086	3.958	catechin	$C_{15}H_{14}O_6$	[M+H]+	93.65
UK -014	417.172	6.971	3-[[4-(2,4-dimethylphenyl) -5-(1-naphthylmethyl) -1,2,4-triazol -3-yl]sulfanyl]propanamide	$C_{24}H_{24}N_4OS$	[M+H]+	59.30
UK -015	275.185	5.841	[4-amino-1-(2-methylpropyl) -6,7,8,9-tetrahydro -1h-imidazo[4,5-c]quinolin -2-yl]methanol	$C_{15}H_{22}N_4O$	[M+H]+	76.53
UK -016	181.072	2.541	4-hydroxy -6-methyl -3,4-dihydropteridin-2(1H)-one	$C_7H_8N_4O_2$	[M+H]+	86.80
UK -017	471.090	8.433	Luteolin 7-b-D-Glucopyranoside	$C_{21}H_{20}O_{11}$	[M+Na]+	84.23
UK -018	565.157	7.027	3,4-dihydroxy-9,10 -dioxo -9,10-dihydroanthracen-2-yl-6-O-(6-deoxy-alpha -L -mannopyranosyl) -beta-D -glucopyranoside	$C_{26}H_{28}O_{14}$	[M+H]+	53.25
UK -019	579.150	4.123	(2r,3s) -2-(3,4-dihydroxyphenyl) -8-{2,3-dihydroxy -5-[(2r,3s) -3,5,7-trihydroxy -3,4-dihydro -2h-chromen-2-yl]phenyl} -3,4-dihydro -2h-chromene-3,5,7-triol	$C_{30}H_{26}O_{12}$	[M+H]+	64.55
UK -020	307.083	1.834	gallocatechin	$C_{15}H_{14}O_7$	[M+H]+	87.86

4. Conclusions

The LCMS-IT-TOF provides positive and negative MS^n data with low variability. In this study, MS^1 data was used to generate a spectrally aligned data array of mass intensity and retention time pairs for multivariate analysis. A PLS regression was performed using this peak set and a tea quality contest ranking as the free and bound variables, respectively. The quality evaluation model was constructed successfully using a PLS regression model. Whole samples except test samples which ranked 10th, 20th and 30th places were used as the training set. We selected 20 features for compound identification which the quality prediction model indicated high importance. Thousands of candidates were initially returned by the database search. By using an automatic workflow involving formula prediction and product ion assignment, the number of candidates were narrowed successfully without non-trivial tasks such as the manual assignment of the product ion spectrum and literature searching. In addition to well-known compounds such as caffeine and catechins, these candidates include various esters of organic acids. This technique is not limited to the analysis of secondary metabolites as reported here, but is also applicable for the prediction of a wide range of compounds, including additives and impurities in polymers and pesticides.

First Edition: June, 2013

SHIMADZU

Shimadzu Corporation
www.shimadzu.com/an/

For Research Use Only. Not for use in diagnostic procedures.
The content of this publication shall not be reproduced, altered or sold for any commercial purpose without the written approval of Shimadzu. The information contained herein is provided to you "as is" without warranty of any kind including without limitation warranties as to its accuracy or completeness. Shimadzu does not assume any responsibility or liability for any damage, whether direct or indirect, relating to the use of this publication. This publication is based upon the information available to Shimadzu on or before the date of publication, and subject to change without notice.