Supplementary Information:

Winners of CASMI2013: Automated Tools and Challenge Data

Takaaki Nishioka[#†1)], Takeshi Kasama[*†2)], Tomoya Kinumi[*†3)], Hidefumi Makabe[*4)], Fumio Matsuda[*5)], Daisuke Miura[*6)], Masahiro Miyashita[*†7)], Takemichi Nakamura[*†8)], Ken Tanaka[*9)], Atsushi Yamamoto[*†10)]

*) These authors equally contributed to the paper as the CASMI2013 organizers.

#) Corresponding author

†) Members of Spectral Data Division, Mass Spectrometry Society of Japan

1) Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan.

2) Research Center for Medical and Dental Sciences, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo 113-8510, Japan.

3) National Metrology Institute of Japan, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568 Japan.

4) Graduate School of Agriculture, Shinshu University, 8304 Minami-minowa, Kami-ina, Nagano 399-4598, Japan.

5) Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565-0871, Japan.

6) Innovation Center for Medical Redox Navigation, Kyushu University, Higashi-ku, Fukuoka, Fukuoka 812-8582, Japan.

7) Graduate School of Agriculture, Kyoto University, Sakyo-ku, Kyoto, Kyoto 606-8502, Japan.

8) Collaboration Promotion Unit, RIKEN Global Research Cluster, Wako, Saitama 351-0198, Japan.

9) College of Pharmaceutical Sciences, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan.

10) Osaka City Institute of Public Health and Environmental Sciences, Tennoji-ku, Osaka, Osaka 543-0026, Japan.

Contents

S1. Methods of candidate preparations

This supplementary section introduces the method that each participant prepared the submitted candidates. The introductions in this section came from the participant's metadata that were attached to their submissions. For more details see the articles submitted by the participant teams to the CASMI2013 special issue of *Mass Spectrometry*.

S1-1. The team of Andrew Newsome and Dejan Nikolic, University of Illinois at Chicago, IL, USA, 'Newsome' team, participated in CASMI2013 with manual method. Abstract of the method is as follows.

Category 1: Formula candidates were determined on a case by case basis using manual methods. The manual methods used to arrive at structural candidates typically involved a combination of monoisotopic fragment ion and neutral loss formula analysis using a formula calculator and Excel spreadsheet as well as database searching on the accurate masses of molecular ions, fragment ions, neutral losses, and potential formulas thereof. The search databases most often employed were ChemSpider, SciFinder Scholar, Reaxys, and Google Scholar. Literature consultation, deductive reasoning, and tacit knowledge and experience were also used. In many cases, the formula could be determined strictly from the accurate mass and fragment ion analysis. In some cases, a formula candidate was not decided upon until after the category 2 structure candidates were determined.

For ranking candidate structures, a subjective confidence scale from 0.60 to 1.00 was used. Structures were placed on the scale based upon how "confident" we felt about the proposed structure from our overall assessment of the fit of the candidates to the challenge data. The confidence scale ranking brackets are defined as follows:

1.00: Full confidence that the single candidate is the correct formula.

0.90 to 0.99: High confidence that candidate is the correct formula.

0.80 to 0.89: Good confidence that candidate is the correct formula.

0.70 to 0.79: Fair confidence that candidate is the correct formula.

0.60 to 0.69: Poor confidence that candidate is the correct formula.

Formula candidates were submitted for all of the challenges. Adduct formulas were provided for challenges 7, 8, 13,and 14. There were no cases where more than one formula was submitted, but some formula submissions were ranked at a higher level of confidence than others.

Category 2: Structure candidates were determined on a case by case basis using manual methods. The manual methods used to arrive at structural candidates typically involved a combination of monoisotopic fragment ion and neutral loss formula analysis, database searching on molecular ion and fragment formulas and monoisotopic masses, literature consultation, deductive reasoning, and tacit knowledge and experience. The search databases most often employed were ChemSpider, SciFinder Scholar, Reaxys, and Google Scholar.

For ranking candidate structures, a subjective confidence scale from 0.60 to 1.00 was used. Structures were placed on the scale based upon how "confident" we felt about the proposed structure from our overall assessment of the fit of the candidates to the challenge data. The confidence scale ranking brackets are defined as follows:

1.00: Full confidence that the single candidate is the correct structure.

0.90 to 0.99: High confidence that candidate is the correct structure.

0.80 to 0.89: Good confidence that candidate is the correct structure.

0.70 to 0.79: Fair confidence that candidate is the correct structure.

0.60 to 0.69: Poor confidence that candidate is the correct structure.

Where several possible structural isomers existed that matched the challenge data, isomers that were thought to be more likely were placed in a higher ranking bracket. In cases where many other possible structures existed that could potentially match the challenge data, we noted this in the respective abstract and lowered the confidence score for the submission accordingly. Structures placed in the same ranking bracket were regarded as equally likely. Structure candidates were submitted for all challenges except for challenge 13.

S1-2. The team of Lars Ridder and Justin J.J. van der Hooft, Wageningen Wageningen University, Laboratory of Biochemistry, Wageningen, The Netherlands and University of Glasgow, College of Medical, Veterinary, and Life Sciences, United Kingdom, ‘Ridder’ team, participated in the contest with an automatic method, MAGMa. Abstract of the method is as follows.

Category 1: The challenge peak lists were converted to MAGMa input files, and processed with MAGMa using candidate molecules from PubChem, as described in the metadata file for category 2. Submissions for category 1 consists of the lists of unique molecular formula's obtained in category 2. The provided scores correspond to the highest scoring candidate (in category 2) with the given molecular formula. This

submission is supported by the observation in Ridder et al. (2012) and Ridder et al. (2013) that, even if the top scoring candidate structure in MAGMa is not correct, the molecular formula often is.

Category 2: The challenge peak lists were converted to MAGMa input files, and processed with MAGMa using candidate molecules from PubChem (Ridder et al. 2012, Ridder et al. 2013). This method is available here: http://www.emetabolomics.org/magma. It does not make use of searches in spectral libraries. By default MAGMa is restricted to candidate molecules from PubChem <1200 Da and consisting of the elements C,H,N,O,P and S. For challenges 7, 8 and 9 candidates molecules were retrieved from PubChem outside the default restrictions. Candidates for 7 and 8 were >1200 Da, and challenge 9 was recognized to contain chlorine atoms, based on the isotope pattern, so candidate molecules with halogens were included. The reported score represents the "refined ranking" as described in Ridder et al. (2013). For challenges 1, 2 and 14 de large numbers of PubChem candidates obtained initially were reduced based on a threshold of 5 on the number of related PubChem references. No submissions are made for challenges 11,12,15 and 16 for which none of the retrieved PubChem candidates (based on default restrictions) provided a satisfactory match in MAGMa between fragment ions and in silico substructures.

References:

L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. van Schaik, J. Vervoort. Substructure-based annotation of high-resolution multistage MSn spectral trees. *Rapid Comm. Mass Spectrom*. 26: 2461-2471, 2012.

L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. J. Bino, J. Vervoort. *Anal. Chem.* 85: 6033-6040, 2013.

S1-3. The team of Kai Dührkop and Sebastian Böcker, Friedrich-Schiller-University, Jena, Germany, 'Dührkop' team, participated in only Category 1 by an automatic method SIRIUS. Dührkop team participated in the first CASMI. Abstract of the method is as follows.

Category 1: The spectral data (MS and MS/MS) was analyzed using the newest (not yet published and still in progress) version of the SIRIUS command line tool. The isotope pattern analysis limits itself to $[M+H]^+$ and $[M+Na]^+$ ions in positive mode and $[M-H]^-$ ion in negative mode. The chosen allowed mass deviation depends on the instrument: Orbitrap: 5 ppm, TOF (positive): 10 ppm, TOF (negative): 20 ppm, FTICR: 2 ppm.

We used the alphabet C, H, N, O, P, S, Cl, Br, I and F but we set upperbounds for certain elements to speed up computations: F, I and S are restricted to 6 occurences per molecule. Cl and P are restricted to 3 occurences per molecule. Br is restricted to one occurence per molecule. For molecules with mass greater than 900 Da we used only the alphabet C, H, N, O, P and S.

The molecular formula identification of SIRIUS is an automatic method. It is complete de-novo and does not perform any database search: Neither in compound databases nor in spectral databases.

The output of SIRIUS is a list of all possible molecular formulas within the allowed mass range together with their scores. We (automatically) transformed this output list to a new representation which is more suitable for this contest:

The best formula candidate gets score 1.0. Following formulas get a logarithmic decreasing score. Formulas which SIRIUS score differs more than 10% from the SIRIUS score of the best candidate formula are excluded.

The challenges 7, 8, 13 and 14 are excluded, as the correct molecular formula is given in the challenge's description. For challenge 15 and 16 we ignored the MS/MS spectra with unit mass resolution.

S1-4. The team of Emma Schymanski and Steffen Neumann, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland and Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany, 'Schymanski' team, participated another participant that submitted the solution candidates prepared by three automatic methods MOLGEN-MS/MS, MetFrag and MetFusion. Schymanski team also participated in the first CASMI. Abstract of the method is as follows.

Category 1: Category 1 challenges were processed with MOLGEN-MS/MS using the elements C, H, N, O, P, and S, where no evidence of halogens was present. Additional parameters were adjusted according to the AnalyticalMethods files. The mode was either $[M+H]^+$ or $[M-H]^-$, depending on whether positive or negative mode was quoted in the files. As most were ESI ionization, this is a reasonable (but not foolproof) assumption.

Some challenges were filtered by ring and double-bond counts as this was given as "clues" (aromatic structure present, amide bonds). Default parameters were MS accuracy 5 ppm, MS/MS accuracy 10 ppm, using the existence filter and allowing "OEI" ions to explain MS/MS peaks.

Generally the combined match value was used as the score, except for challenge 15 where no MS isotope pattern was present - in this case MSMSMV scaled by ppm was used.

Where multiple MS/MS files were available, these were scaled to relative intensities and merged taking the peak of highest relative intensity where the same peak occurred more than once. Isotope patterns in the MS/MS were removed where the accuracy in the MS/MS was sufficient to unequivocally identify the peak as an isotope and not a fragment. The results were cross-checked with the category 2 submissions.

Category 2: Category 2 challenges were processed using MetFrag and MetFusion using compound database queries. Three databases (PubChem, ChemSpider and KEGG) were queried and the results were merged to create one candidate list, taking the maximum score for entries in more than one database. MetFusion also used MassBank.jp to retrieve spectral information, with default parameters. Information from the Analytical Methods files and Category 1 results (ES) were used to adjust input parameters for the automatic pipeline.

The formula was used for candidate retrieval where this was given or clear from Category 1, otherwise the exact mass was used. We enabled the element filter C, H, N, O, P, and S unless we obtained high scoring non-C, H, N, O, P, and S candidates without it. The exact mass database retrieval and fragmentation parameters were adjusted according to the expected or quoted instrument accuracy. Manual checking of the automatic calculations were performed to detect any anomalies.

Where the MetFusion scores were poor (few or no matching spectra from MassBank), MetFrag results were submitted.

S1-5. Daniel L. Sweeney, MathSpec, Inc., IL, USA, 'Sweeney', participated in only the Category 2 challenges by commercially available automatic methods and manual method. Abstract of the method is as follows.

Category 2: Used Rational Numbers Search software and searched the mass spectral peak lists against a database of approximately 200,000 molecules that had been partitioned for rapid mass spectral searching. The InChI structures were copied from the PubChem entry for the corresponding compound.

In challenges where the precursor ion was not present in the MS/MS data file, the precursor ion was copied from the MS data file. All ions greater in mass than the

precursor ion, if present in the MS/MS data file, were removed prior to the analysis. The isotope data was not used.

Challenges 7 and 8 were done manually. Challenge 10 was done manually with the aid of an Excel Add-In.

Attempts were made to identify all sixteen compounds, but no possibilities were found for challenges 3 or 13.

S1-6. The team of Felicity Allen and Russ Greiner, University of Alberta, Alberta, Canada, 'Allen' team, participated in Category 1 and 2 challenges by an automatic method, CFM. The team missed an opportunity to submit their paper to the special issue of Mass Spectrometry. Abstract of the method follows.

A list of candidate structures was obtained by querying PubChem for all structures within 10ppm of the precursor mass (or with the given molecular formula if this was provided). For cases where the precursor mass was not provided, this value was deduced manually by considering the MS2 and MS1 data. Where further specific information was provided, the candidate lists were filtered using that information e.g. aromaticity, amide bonds.

The candidate lists were then processed with the input spectra by the program cfm-id (http://sourceforge.net/projects/cfm-id/) to produce a ranked list of structures for Category 2. A Single-Energy CFM model was used, for which parameters were trained using non-peptide metabolite data from METLIN, as described in http://arxiv.org/abs/1312.0264 and stored in the supplementary data section of the above sourceforge project. Since the model expects a low, medium and high energy spectrum, whereas the challenge data (except 16) only has one spectrum, we repeated the provided spectrum for all three energy levels. For Challenge 16, we repeated the CE20 spectrum for low and medium and used the CE40 spectrum for high. All spectra were pre-processed - peaks below 1% intensity relative to the highest peaks were removed.

For Category 1, the molecular formula was computed for each structure from Category 2 and kept in the same order. The list was then processed to remove duplicate entries, keeping only the highest ranked listing for each unique molecular formula.

Submission is only made for positive ion mode, since cfm-id does not currently support negative mode.

S1-7. Tsubasa Miyazaki and Hisayuki Horai, Ibaraki National College of Technology, Ibaraki, Japan, participated with the candidates manually prepared and resulted in no correct candidate to the Category 2 challenges. The metadata of this team was written by unclear English.

Table S1. Details per Challenge and Participant in Category 1. See the Table legend at the bottom for more details.

Participants	Challenges	rank	tc	bc	ec
Newsome	challenge1	1	1	0	1
Schymanski	challenge1	2	11	0	2
Allen	challenge1	1	34	0	1
Dührkop	challenge1	1	1	0	1
Ridder	challenge1	1	5	0	1
Newsome	challenge2	1	1	0	1
Schymanski	challenge2	1	2	0	1
Allen	challenge2	-	5	-	-
Dührkop	challenge2	1	1	0	1
Ridder	challenge2	1	2	0	1
Newsome	challenge3	1	1	0	1
Schymanski	challenge3	1	3	0	1
Allen	challenge3	-	4	-	-
Dührkop	challenge3	1	1	0	1
Ridder	challenge3	1	5	0	1
Newsome	challenge4	1	1	0	1
Schymanski	challenge4	1	4	0	1
Allen	challenge4	1	13	0	1
Dührkop	challenge4	1	1	0	1
Ridder	challenge4	1	2	0	1
Newsome	challenge5	1	1	0	1
Schymanski	challenge5	1	9	0	1
Allen	challenge5	3	37	2	1
Dührkop	challenge5	1	1	0	1
Ridder	challenge5	1	4	0	1

Newsome	challenge6	1	1	0	1
Schymanski	challenge6	8	144	7	1
Dührkop	challenge6	1	1	0	1
Ridder	challenge6	1	2	0	1
Newsome	challenge9	1	1	0	1
Schymanski	challenge9	1	6	0	1
Allen	challenge9	1	21	0	1
Dührkop	challenge9	1	1	0	1
Ridder	challenge9	1	15	0	1
Newsome	challenge10	1	1	0	1
Schymanski	challenge10	45	283	44	1
Allen	challenge10	1	10	0	1
Dührkop	challenge10	1	18	0	1
Ridder	challenge10	1	7	0	1
Newsome	challenge11	1	1	0	1
Schymanski	challenge11	4	41	3	1
Dührkop	challenge11	1	1	0	1
Newsome	challenge12	1	1	0	1
Schymanski	challenge12	9	27	8	1
Dührkop	challenge12	-	1	-	-
Newsome	challenge15	1	1	0	1
Schymanski	challenge15	-	177	-	-
Dührkop	challenge15	-	11	-	-
Newsome	challenge16	1	1	0	1
Schymanski	challenge16	-	6	-	-
Allen	challenge16	2	51	1	1
Dührkop	challenge16	1	1	0	1

Table legend:

Newsome, Schymanski, Allen, Dührkop and Ridder are team names.
"-" and blank show that challenge has a submission with no correct candidate and no submission, respectively.
rank: Absolute rank of correct candidate defined by Equation 1 (see Text).
tc: Total number of candidates submitted.
bc: Number of candidates with a score better than the correct candidate.
ec: Number of candidates with the same score as the correct candidate.

Table S2. Details per Challenge and Participant in Category 2. See the Table legend at the bottom for more details.

Participant	Challenge	rank	tc	bc	ec
Newsome	challenge1	1	2	0	1
Sweeney	challenge1	1	1	0	1
Schymanski	challenge1	9	5631	8	1
Allen	challenge1	12	6767	9	3
Ridder	challenge1	1	1084	0	1
Miyazaki	challenge1	-	1	-	-
Newsome	challenge2	1	2	0	1
Sweeney	challenge2	1	1	0	1
Schymanski	challenge2	44	12702	43	1
Allen	challenge2	-	131	-	-
Ridder	challenge2	3	631	2	1
Miyazaki	challenge2	-	1	-	-
Newsome	challenge3	-	1	-	-
	challenge3 (Ile)	1	1	0	1
Schymanski	challenge3	21	335	0	21
	challenge3 (Ile)	21	335	0	21
Allen	challenge3	-	18	-	-
Ridder	challenge3	17	370	2	15
	challenge3 (Ile)	2	370	0	2
Miyazaki	challenge3	-	1	-	-
Newsome	challenge4	1	1	0	1
Sweeney	challenge4	-	10	-	-
Schymanski	challenge4	238	721	236	2
	challenge4 (4mp)	299	721	298	1
	challenge4 (2mp)	293	721	292	1
Allen	challenge4	18	1622	16	2
	challenge4 (4mp)	4	1622	0	4
	challenge4 (2mp)	4	1622	0	4
Ridder	challenge4	78	825	77	1
	challenge4 (4mp)	75	825	74	1
	challenge4 (2mp)	76	825	75	1

Miyazaki	challenge4	-	1	-	-
Newsome	challenge5	1	3	0	1
	challenge5 (propyl)	2	3	1	1
Sweeney	challenge5	1	2	0	1
Schymanski	challenge5	4	366	3	1
	challenge5 (propyl)	1	366	0	1
Allen	challenge5	9	2725	8	1
	challenge5 (propyl)	42	2725	40	2
Ridder	challenge5	2	350	1	1
	challenge5 (propyl)	1	350	0	1
Miyazaki	challenge5	-	1	-	-
Newsome	challenge6	1	1	0	1
Sweeney	challenge6	1	1	0	1
Schymanski	challenge6	1	6	0	1
Ridder	challenge6	1	2	0	1
Miyazaki	challenge6	-	1	-	-
Newsome	challenge7	1	7	0	1
Sweeney	challenge7	1	1	0	1
Schymanski	challenge7	17	17	0	17
Allen	challenge7	23	24	14	9
Ridder	challenge7	1	17	0	1
Miyazaki	challenge7	-	1	-	-
Newsome	challenge8	1	3	0	1
Sweeney	challenge8	2	2	0	2
Schymanski	challenge8	1	1	0	1
Allen	challenge8	1	1	0	1
Ridder	challenge8	1	1	0	1
Miyazaki	challenge8	-	2	-	-
Newsome	challenge9	1	6	0	1
Sweeney	challenge9	1	1	0	1
Schymanski	challenge9	1	4	0	1
Allen	challenge9	2	150	1	1
Ridder	challenge9	1	113	0	1
Miyazaki	challenge9	-	1	-	-
Newsome	challenge10	1	2	0	1

Sweeney	challenge10	1	1	0	1
Schymanski	challenge10	1	9	0	1
Allen	challenge10	1	20	0	1
Ridder	challenge10	1	20	0	1
Miyazaki	challenge10	-	3	-	-
Newsome	challenge11	2	3	1	1
	challenge11 (tautomer1)	3	3	2	1
	challenge11 (tautomer2)	1	3	0	1
Sweeney	challenge11	6	17	5	1
	challenge11 (tautomer1)	5	17	4	1
	challenge11 (tautomer2)	-	17	-	-
Schymanski	challenge11	21	2392	20	1
	challenge11 (tautomer1)	1	2392	0	1
	challenge11 (tautomer2)	22	2392	21	1
Newsome	challenge12	1	3	0	1
Sweeney	challenge12	3	21	2	1
Schymanski	challenge12	35	902	34	1
Schymanski	challenge13	12	227	11	1
Allen	challenge13	24	284	18	6
Ridder	challenge13	42	206	41	1
Newsome	challenge14	1	1	0	1
Sweeney	challenge14	2	5	1	1
Schymanski	challenge14	1	8219	0	1
Allen	challenge14	761	9708	732	29
Ridder	challenge14	5	1583	4	1
Newsome	challenge15	1	3	0	1
Sweeney	challenge15	1	4	0	1
Schymanski	challenge15	-	6	-	-
Newsome	challenge16	1	1	0	1
Sweeney	challenge16	1	4	0	1
Schymanski	challenge16	-	3976	-	-
Allen	challenge16	100	10637	97	3

Table legend:
Newsome, Sweeney, Schymanski, Allen, Ridder and Miyazaki are team names.

rank: Absolute rank of correct candidate defined by Equation 1 (see Text).
tc: Total number of candidates submitted
bc: Number of candidates with a score better than the correct candidate
ec: Number of candidates with the same score as the correct candidate