Critical Assessment of Small Molecule Identification: Results of the Inaugural CASMI Contest 2012

Emma Schymanski (1), Steffen Neumann (2)

(1) Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, Dübendorf, 8600, Switzerland; emma.schymanski@eawag.ch
(2) Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle (Saale), Germany; sneumann@ipb-halle.de

Aims

CASMI

Critical Assessment of Small Molecule Identification

Exchange ideas and strategies for identifying unknowns between disciplines by:

- *Providing* a common MS dataset of unknowns
- *Allowing* a systematic, comparative evaluation of different (automated or human) methods

Contest website:
`http://www.casmi-contest.org/`

Contest Categories

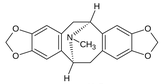
Four categories with two MS technologies:

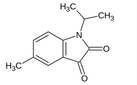
- Category 1: Best molecular *formula*, LC-HRMS/MS
- Category 2: Best molecular *structure*, LC-HRMS/MS
- Category 3: Best molecular *formula*, GC-MS
- Category 4: Best molecular *structure*, GC-MS

Challenges and Solutions

30 challenges in total:

- Plant material (e.g. top right)
- Environmental samples
- Reference standards (e.g. bottom right)





Categories 1 & 2: 14 challenges:

- 6 microTOF ESI MS/MS
- 8 Orbitrap ESI or APCI MS/MS
- Ions included $[M+H]^+$, $[M-H]^-$ and some tricky cases

Categories 3 & 4: 16 challenges (GC-MS)

Participation

Eight participants submitted entries:

- A total of 130 challenge submissions
- Categories 1 and 2 were most popular

Submissions

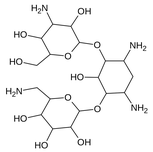
Entries submitted as tab-separated files with:

- Molecular formulas (Categories 1 and 3)
- SMILES or InChI structures (Categories 2 and 4)
- Scores for each candidate formula/structure
  - the higher the score the better!

Evaluation of submissions was *fully automatic*
`http://casmi-contest.org/rules.shtml`

Example Submissions Category 1

Challenge 1: Kanamycin A:



(a) Dunn *et al.*

Candidate	Score
$C_{18}H_{36}N_4O_{11}$	1.00

(b) Dührkop *et al.*

Candidate	Score
$C_{19}H_{32}N_8O_7$	1.36e+16
$C_{22}H_{38}N_4O_4P_2$	5.20e+13
$C_{18}H_{36}N_4O_{11}$	1.46e+13
$C_{17}H_{30}N_{11}O_6$	1.49e+12
$C_{20}H_{43}N_2O_5P_3$	1.08e+11
$C_{20}H_{41}N_2O_7PS$	8.02e+10

(c) Shen *et al.*

Candidate	Score
$C_{30}H_{44}O_5$	0.89
$C_{23}H_{44}N_6O_5$	0.58
$C_{25}H_{33}O_8$	0.56
$C_{18}H_{36}N_4O_{11}$	0.20
$C_{29}H_{40}O_6$	0.15

Automatic Evaluation Measures

Several evaluation measures were used:
(BC, EC, WC, TC = better, equal, worse, total candidates)

- Absolute Rank: $Rank_{WorstCase} = BC + EC$
- Relative Rank: $RRP = \frac{1}{2}\left(1 - \frac{BC-WC}{TC-1}\right)$
- Normalised Score: $\tilde{s}_j = \frac{s_j}{\sum_i s_i}$ such that $\sum_i \tilde{s}_i = 1$
- Weighted Relative Rank: $wRRP = 1 - \sum_{\tilde{s}_i > \tilde{s}_j} \tilde{s}_i - \sum_{\tilde{s}_i = \tilde{s}_j} \tilde{s}_i$
- Structural Similarity: Tanimoto similarity (TS) using CDK fingerprints and Tanimoto distance

[301 638 742 743 930]; [638 743]; $TS = 0.4$

The **absolute rank** was used to *declare the winners of CASMI 2012!*

Category 1 - Best Molecular Formula

. . . and the LC-MS/MS winner is: Dunn *et al.*

Participant Name	Entries Subm.	Correct (wins)	Avg. Rank	Avg. TC	Avg. RRP
Dunn *et al.*	**11**	**9(9)**	**1.11**	**1.4**	**0.500**
Shen *et al.*	14	8(3)	2.88	11.1	0.719
Dührkop *et al.*	14	12(5)	5.75	134.7	0.883
Neumann *et al.*	14	12(*NA*)	4.83	1915	0.997
Meringer *et al.*	14	14(*NA*)	3.29	36.6	0.941

- **MS, isotopes, MS/MS and manual interpretation yielded best molecular formula results**

Challenge 3 and Submissions

Automatic spectrum annotation via SMILES, OpenBabel and R.

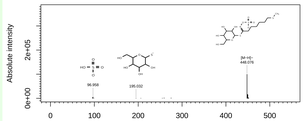


Figure 1a: Annotated, merged ESI negative MS/MS spectra of glucolesquerellin

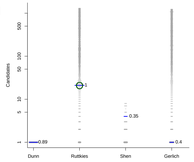


Figure 1b: Similarity calculations for Challenge 3 submissions

- Huge variety in number and closeness of entries
- Fig. 1b: two participants had wrong molecular formula
- Similarity allows assessment of "closeness"

Challenge 3: Most Similar Entries

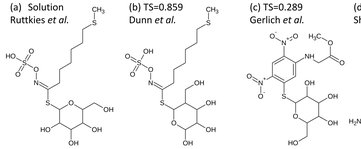


Figure 2: Most similar entries for Challenge 3.

- Fig. 2b: substitution isomer of correct compound
- Fig. 2c&d: entries had the incorrect formula and thus different functional groups

Challenge 10 and Submissions

Automatic annotation of SMILES via *in silico* fragmentation methods.

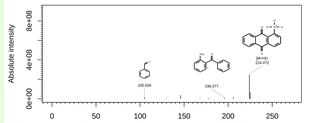


Figure 3a: Annotated, merged ESI positive MS/MS spectra of 1-aminoanthraquinone

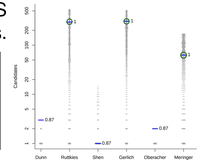


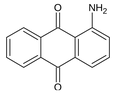
Figure 3b: Similarity calculations for Challenge 10 submissions

Look at Figure 4 - which scenario is "more correct"?

- Correct answer at Rank ≈ 60 or wrong isomer at Rank 1?

Challenge 10: Most Similar Entries

(a) Solution
Ruttkies *et al.*
Gerlich *et al.*
Meringer *et al.*



(b) TS=0.842
Dunn *et al.*
Shen *et al.*
Oberacher

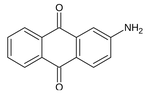


Figure 4: Most similar entries for Challenge 10.

- Most similar entry is a substitution isomer; present in KEGG, MassBank
- Correct identification very dependent on database

Category 2 - Best Molecular Structure

. . . and the LC-MS/MS winner is: Oberacher

Participant Name	Entries Subm.	Correct (wins)	Avg. Rank	Avg. TC	Avg. RRP
Dunn *et al.*	11	3(2)	5.7	4.7	0.556
Ruttkies *et al.*	14	14	319.7	1226.3	0.838
Shen *et al.*	14	4(2)	5.5	19.4	0.614
Gerlich *et al.*	14	14	305.4	2878.1	0.873
Oberacher	**5**	**3(3)**	**1.0**	**1.2**	**—**
Meringer *et al.*	6	6	29.2	258.5	0.551

- **Spectral libraries best for molecule identification!**
- **Mixed success with compound databases: lower average ranks; more misses with smaller databases**
- **More spectra are needed for public libraries!**

CASMI Proceedings

Special issue of the MDPI Open Access journal *Metabolites*

- Schymanski, E.L.; Neumann, S.
  CASMI: Challenges and solutions.
- Allwood, J.W.; Weber, R.J.M.; Zhou, J.; He, S.; Viant, M.; Dunn, W.B.
  CASMI-the small molecule identification process from a Birmingham perspective.
- Shen, H.; Zamboni, N.; Heinonen, M.; Rousu, J.
  Metabolite identification through machine learning-tackling CASMI challenges using FingerID.
- Dührkop, K.; Scheubert, K.; Böcker, S.
  Molecular formula identification with SIRIUS.
- Oberacher, H.
  Applying tandem mass spectral libraries for solving the CASMI LC/MS challenge 2012.
- Ruttkies, C.; Gerlich, M.; Neumann, S.
  Tackling challenging challenges with MetFrag and MetFusion.
- Meringer, M.; Schymanski, E.L.
  Small molecule identification with MOLGEN and mass spectrometry.
- Schymanski, E.L.; Neumann, S.
  CASMI: And the Winner is . . .

`www.mdpi.com/journal/metabolites/special_issues/CASMI`

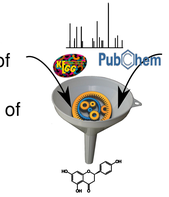
CASMI Continues . . .

Look out for CASMI 2013/2014

- Coordinator: Prof. Takaaki Nishioka
- Team: Mass Spectrometry Society of Japan (MSSJ).

Participants welcome from all fields of mass spectrometry

- Automated approaches *through to* expert, manual data evaluation



Acknowledgements

Thanks to those who contributed data, the board members and participants. ES was funded by the European Union, Marie Curie Postdoctoral Fellowship Grant PIEF-GA-2011-299734. No financial gains were made for organising or participating in CASMI.