A new approach to identifying unknown trace level analytes by tandem mass spectrometry without reference spectroscopic database support: CSI:FingerID

Antony N. Davies[a,b]
[a]Strategic Research Group – Measurement and Analytical Science, Akzo Nobel, Deventer, the Netherlands
[b]SERC, Sustainable Environment Research Centre, Faculty of Computing, Engineering and Science, University of South Wales, UK

Tandem mass spectrometry[1] has become a tool of choice for the identification of trace levels of compounds, especially in data-rich environments such as metabolomics and increasingly for screening in toxicological or forensic investigations. The benefits of adopting such methods lie in the low detection limits and the ability to extract quantitative information from samples with impossibly difficult to impossible clean-up options with multiple analytes co-eluting from the chromatography stage. Most of these techniques rely heavily on prior knowledge of the ionisation profiles of the analytes being studied. Where the metabolites are potentially unknown, it would be possible to resort to database searching of the signals observed, however, here, as in many areas of spectroscopy, the lack of comprehensive reference database coverage means that the spectroscopist has to resort to other methods to identify the newly detected compounds. This article looks at the background to the problem and a novel solution—CSI:FingerID—which allows for highly sensitive tandem mass *spectrometry* (MS/MS) data to be used to identify unknown analytes from common molecular structure databases where reference spectroscopic data is unavailable.

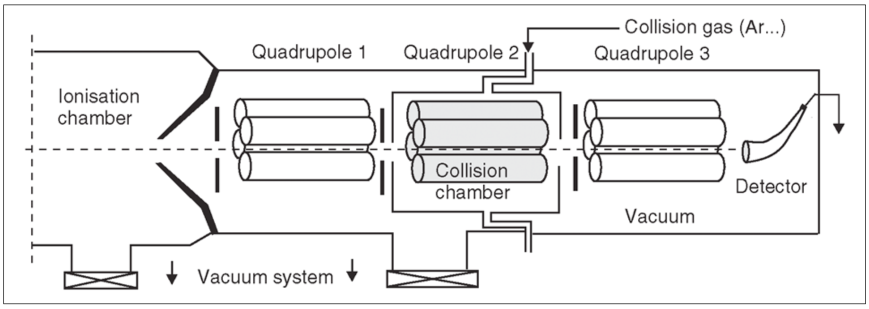


Figure 1. A schematic of a simple triple quadrupole tandem mass spectrometer. The middle quadrupole is used as a collision chamber. These instruments can conduct all three common types of MS/MS analysis shown in Figure 2. For the untargetted metabolomics experiment a much higher mass resolution would be required such as from a Q-ToF or Orbitrap. Reproduced from Reference 2 with permission. © John Wiley & Sons.

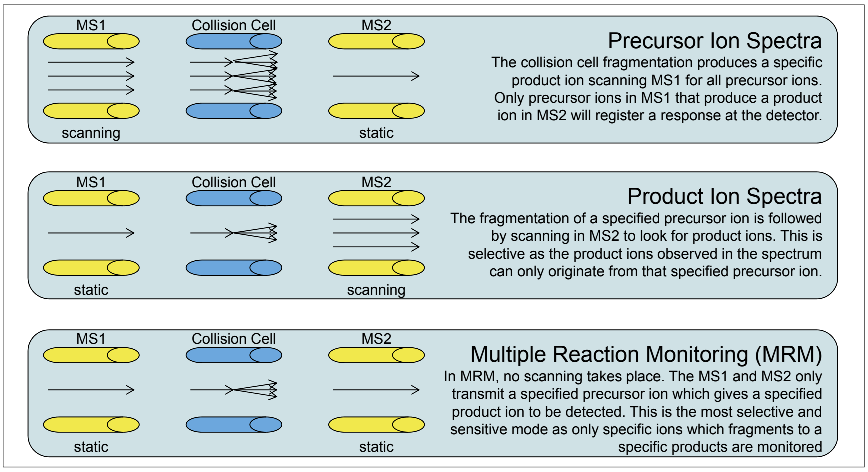

MS1
Collision Cell
MS2
scanning
static
Precursor Ion Spectra
The collision cell fragmentation produces a specific product ion scanning MS1 for all precursor ions. Only precursor ions in MS1 that produce a product ion in MS2 will register a response at the detector.
MS1
Collision Cell
MS2
static
scanning
Product Ion Spectra
The fragmentation of a specified precursor ion is followed by scanning in MS2 to look for product ions. This is selective as the product ions observed in the spectrum can only originate from that specified precursor ion.
MS1
Collision Cell
MS2
static
static
Multiple Reaction Monitoring (MRM)
In MRM, no scanning takes place. The MS1 and MS2 only transmit a specified precursor ion which gives a specified product ion to be detected. This is the most selective and sensitive mode as only specific ions which fragments to a specific products are monitored

Figure 2. Three of the most common MS/MS experiments following the initial ionisation step. 1) Scanning the first quadrupole whilst keeping the second fixed on a specific product ion; 2) keeping the first quadruple fixed on a specific precursor ion whilst scanning the second to obtain a full product ion spectrum; or 3) keeping both quadrupoles fixed to only allow a specific precursor and product ion pair to be detected.

Tandem mass spectrometry

So, what are we talking about? Tandem mass spectrometry has been available for many years and deployed in many fields where identification and quantification of analytes is not possible by a single mass detector coupled to a chromatography sample preparation step. There are a few common main instrumental configurations depending on experiments to be performed. Figure 1 from Rouessac and Rouessac[2] provides a common instrumental schematic for an instrument capable of MS/MS experiments using three quadrupoles.

Targeted vs untargeted tandem mass spectrometry

Clearly, some of the experiments indicated in Figure 2 require prior knowledge of exactly which analytes you wish to detect and their primary and secondary ionising fingerprints. These "targeted" analyses can be extremely sensitive and selective for specific analytes such as in screening for banned substances in a forensic environment or toxicological screening. This is less useful in areas such as metabolomics studies of new biological samples, where the vast majority of the metabolites observed are initially unknown. Gary Patti and co-workers[3] have a simple to understand figure in their paper showing how the targeted and untargeted approaches play different roles in metabolomics—targeted answering questions around specific levels of known metabolites in samples with untargeted looking at the global metabolic profile of the sample and showing the complex MS/MS picture being the basis for subsequent measurement of MS/MS standards to support substance identification.

CSI:FingerID supporting interpretation of untargeted MS/MS analyses

Sebastian Böcker from the Friedrich Schiller University in Jena, Germany, with his colleagues Kai Dührkop and Marvin Meusel along with Juho Rousu and Huibin Shen from the Institute for Information Technology at Aalto University in Espoo, Finland, looked at the problem of structural elucidation of the results of untargeted tandem mass spectrometry experiments caused by the lack of good quality reference MS/MS data. The approach described by Patti—although eventually providing good results—relies on the availability of reference standard materials for the individual metabolites if it is not to be time consuming. The failure of the community to generate large, good quality reference data collections is compounded by the enormous improvements in instrumentation and sample preparation techniques delivered by the instrument manufacturers. It does provide, however, an excellent hunting ground for innovative bioinformaticians.

The current CSI:FingerID has been built on experiences gained from previous work[4] in this area and has three main sections. Before looking at the unknown spectra, the system is trained in a learning phase—calculating fragmentation trees from known reference spectra, fragmentation tree similarities as well as PubChem (CACTVS)[5] and Klekota–Roth fingerprints[6] (see Table 1). In the current version, each molecular property is predicted by an individual "support vector machine" (SVM) which together make up the nerve centre for the approach. The SVMs yield probabilities for the presence or absence of a particular molecular property in any given compound based on the MS/MS spectral data.

Table 1. Data sources and sizes.

Data source	Number	Use
GNPS Public Spectral Libraries[6]	4138	Training spectra
	3868	Validation spectra
MassHunter Forensics/Toxicology PCDL library[7]	2120	Training spectra
	2055	Validation spectra
MassBank[8]	625	Validation spectra
PubChem[5]	52,926,405 40,805,940	Compounds Structures
PubChem filtered biodatabase[9]	~300,000 268,633	Compounds Structures

With the system trained, it is now possible to move to the actual analysis of new data. The system takes one or more MS/MS spectra, the unknown analyte to be identified and calculates the similarities shown in the unknown data set against the MS/MS spectra and fragmentation trees in the training set. The SVMs then predict the presence or absence of all molecular properties for the unknown compound providing a probability fingerprint.

This fingerprint can then be used as a search criterion against the much bigger molecular structure databases such as PubChem. Each potential solution chemical structure has its calculated fingerprint scored against the unknown to provide a hit-list to the user.

CSI:FingerID validation and comparison against existing methods

For the validation of this approach against existing prediction algorithms, the authors took the normal step of ensuring that none of the compounds used for testing were in the phase one training data sets. Ten-fold validation was carried out ensuring that no repeat batches contained the same structures. The authors provide a lot of statistics for different tests they have carried out against earlier versions of their software and other approaches for this task. Suffice to say that the latest version has yielded a 2.5-times improvement in predicting the correct molecular structure for the MS/MS unknown to 34.4% in the first place in their results list against the PubChem database and an impressive 63.5% of the unknown MS/MS spectra had their correct structures predicted in the top five hits. This is a very strong result.[9]

As phase one uses a machine learning stage which includes not only fragmentation tree information but also reference, known MS/MS data sets to train the system, the authors have also attempted to quantify how this size of this training data set alters the prediction capacity for the system as a whole. By reducing the amount of reference data, they could degrade their prediction capabilities, and carrying out a series of experiments came to the unsurprising conclusion that, with the limited amount of available good quality reference MS/MS data, they were far from saturating the prediction capability of their algorithms, the quality roughly increasing by around one percentage point for each additional 400 reference data sets added to the system.

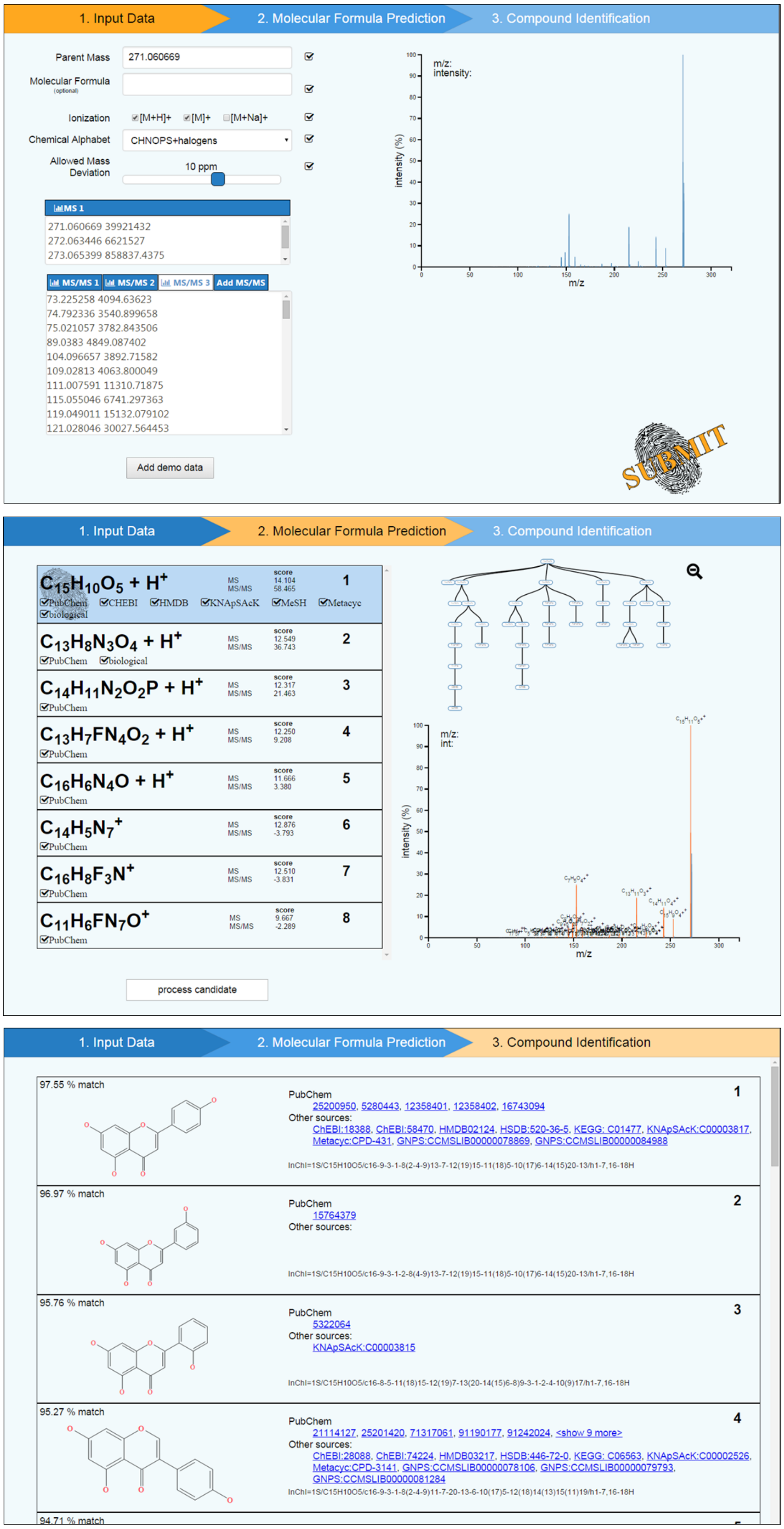
So—a clear call for more data to be made available—again!

Try it out!

CSI:FingerID is available for you to try out at http://www.csi-fingerid.org/. There is a simple user interface (Figures 3–5) and what I really like the ability to run through the sequences of steps using demo data which quickly allows the user to see what they should be inputting.

References

1. IUPAC, *Compendium of Chemical Terminology*, 2nd Edn (the "Gold Book"), Compiled by A.D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997); M. Nic, J. Jirat and B. Kosata, XML on-line corrected version, updates compiled by A. Jenkins (2006). doi: http://dx.doi.org/10.1351/goldbook
2. F. Rouessac and A. Rouessac, *Chemical Analysis, Modern Instrumentation Methods and Techniques*, 2nd Edn. John Wiley & Sons Ltd, Fig. 16.5, p. 401 (2007). ISBN 978-0-470-85903-2
3. G.J. Patti, O. Yanes and G. Siuzdak, "Innovation: Metabolomics: the apogee of the omics trilogy", *Nat. Rev. Mol. Cell Biol.* **13,** 263–269 (2012). doi: http://dx.doi.org/10.1038/nrm3314
4. S. Böcker and F. Rasche, "Towards *de novo* identification of metabolites by analysing tandem mass spectra", *Bioinformatics* **24(16),** i49–i55 (2008). doi: http://dx.doi.org/10.1093/bioinformatics/btn270
5. https://pubchem.ncbi.nlm.nih.gov/
6. https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp
7. Agilent Technologies Inc.
8. H. Horai *et al.*, "MASSBANK: a public repository for sharing mass spectral data for life sciences", *J. Mass Spectrom.* **45(7),** 703–714 (2010). doi: http://dx.doi.org/10.1002/jms.1777
9. K. Dührkop, H. Shen, M. Meusel, J. Rousu and S. Böcker, "Searching molecular structure databases with tandem mass spectra using CSI:FingerID", *Proc. Natl. Acad. Sci. USA* in press. doi: http://dx.doi.org/10.1073/pnas.1509788112



Parent Mass 271.060669 ☑

Molecular Formula (optional) ☑

Ionization ☑[M+H]+ ☑[M]+ □[M+Na]+ ☑

Chemical Alphabet CHNOPS+halogens ☑

Allowed Mass Deviation 10 ppm ☑

MS 1

271.060669 39921432
272.063446 6621527
273.065399 858837.4375

MS/MS 1 MS/MS 2 MS/MS 3 Add MS/MS

73.225258 4094.63623
74.792336 3540.899658
75.021057 3782.843506
89.0383 4849.087402
104.096657 3892.71582
109.02813 4063.800049
111.007591 11310.71875
115.055046 6741.297363
119.049011 15132.079102
121.028046 30027.564453

$C_{15}H_{10}O_5 + H^+$ MS MS/MS score 14.104 58.465 1
☑PubChem ☑CHEBI ☑HMDB ☑KNApSAcK ☑MeSH ☑Metacyc ☑biological

$C_{13}H_8N_3O_4 + H^+$ MS MS/MS score 12.549 36.743 2
☑PubChem ☑biological

$C_{14}H_{11}N_2O_2P + H^+$ MS MS/MS score 12.317 21.463 3
☑PubChem

$C_{13}H_7FN_4O_2 + H^+$ MS MS/MS score 12.250 9.208 4
☑PubChem

$C_{16}H_6N_4O + H^+$ MS MS/MS score 11.666 3.380 5
☑PubChem

$C_{14}H_5N_7^+$ MS MS/MS score 12.876 -3.793 6
☑PubChem

$C_{16}H_8F_3N^+$ MS MS/MS score 12.510 -3.831 7
☑PubChem

$C_{11}H_6FN_7O^+$ MS MS/MS score 9.667 -2.289 8
☑PubChem

97.55 % match 1
PubChem
25200950, 5280443, 12358401, 12358402, 16743094
Other sources:
ChEBI:18388, ChEBI:58470, HMDB02124, HSDB:520-36-5, KEGG: C01477, KNApSAcK:C00003817, Metacyc:CPD-431, GNPS:CCMSLIB00000078869, GNPS:CCMSLIB00000084988
InChI=1S/C15H10O5/c16-9-3-1-8(2-4-9)13-7-12(19)15-11(18)5-10(17)6-14(15)20-13/h1-7,16-18H

96.97 % match 2
PubChem
15764379
Other sources:
InChI=1S/C15H10O5/c16-9-3-1-2-8(4-9)13-7-12(19)15-11(18)5-10(17)6-14(15)20-13/h1-7,16-18H

95.76 % match 3
PubChem
5322064
Other sources:
KNApSAcK:C00003815
InChI=1S/C15H10O5/c16-8-5-11(18)15-12(19)7-13(20-14(15)6-8)9-3-1-2-4-10(9)17/h1-7,16-18H

95.27 % match 4
PubChem
21114127, 25201420, 71317061, 91190177, 91242024, <show 9 more>
Other sources:
ChEBI:28088, ChEBI:74224, HMDB03217, HSDB:446-72-0, KEGG: C06563, KNApSAcK:C00002526, Metacyc:CPD-3141, GNPS:CCMSLIB00000078106, GNPS:CCMSLIB00000079793, GNPS:CCMSLIB00000081284
InChI=1S/C15H10O5/c16-9-3-1-8(2-4-9)11-7-20-13-6-10(17)5-12(18)14(13)15(11)19/h1-7,16-18H

94.71 % match

Figures 3-5. Top: First stage—input your data. Middle: Second stage—the prediction engine works on the possible molecular formulae. Bottom: Third stage—Potential compounds are identified from the PubChem data collection.