MANCHESTER 1824
The University of Manchester

The University of Manchester Research

SIMS informatics

DOI:
10.1002/sia.5065

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Henderson, A., Moore, J. D., & Vickerman, J. C. (2013). SIMS informatics. *Surface and Interface Analysis*, *45*(1), 471-474. https://doi.org/10.1002/sia.5065

Published in:
Surface and Interface Analysis

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Download date:08. Jun. 2022

SIMSInformatics

Alex Henderson[1,2,*], Jimmy D. Moore[1] and John C. Vickerman[1,2]

[1]Manchester Interdisciplinary Biocentre, University of Manchester, Oxford Road, Manchester, M13 9PL, UK

[2]SurfaceSpectra Ltd., Manchester, UK

[*]Corresponding author email: alex.henderson@manchester.ac.uk

Abstract

SIMS instrumentation has improved dramatically over the last decade; however our understanding of the data is yet to catch up. High mass resolution coupled with enhanced transmission and increases in duty cycle have revealed subtle detail in spectral features indicating that there is a wealth of information in our data that we have yet to take advantage of. Here we present a collection of approaches that should improve our data interpretation. We cover websites delivering LC-MS/MS data as web services, allowing us to interact with collections of spectra produced via a similar fragmentation process. We show how simple spectral matching approaches suggest that we can compare data across disparate primary ion sources and mass analyser types, and we show how the management and use of metadata will help us to build resources to the benefit of future researchers.

The role of the SIMS analyst is already complicated by immediacy of instrumental concerns and experimental issues and it is easy for data analysis, archiving and mining to be given reduced attention. In this paper we highlight approaches, methods and tools that can make an immediate impact on the way we view our data and how, with some foresight now, we may be able to make the most of it in the future.

Resources

Collections

When trying to interpret the spectrum of an unknown, there is little that can compete with spectra of standard samples analysed on the users own instrumentation. Here, any anomalous instrumental artefacts or experimental procedure will be present in both the ‘standard’ sample and the unknown. Following the round-robin study coordinated by NPL under the auspices of the VAMAS organisation, the typical error in spectral comparison was found to be better than 10%[1]. This then allows for the use of standard spectra analysed in other careful laboratories to be used in addition to the users’ own. Instrument manufacturers often bundle libraries of standard spectra with their instruments and there are third-party databases such as The SurfaceSpectra Static SIMS Library[2] commercially available.

Given the enormous number of possible compounds that could be present in a sample, notwithstanding the analyst's knowledge of the sample domain, it is impossible to generate a comprehensive resource of spectral standards. ChemSpider has chemical structure information for over 26 million molecules[3] while the CAS Registry recently reported over 60 million compound registrations[4]. Therefore the analyst must resort to other means to interpret their data.

With the ever increasing mass resolution of the modern SIMS spectrometer, one such approach is to determine the molecular formula of a given spectral feature. In 2006, Kind and Fiehn released a database of all 'correct' molecular formulae of mass less than 500 u consisting of C, H, O, N, S and P[5]. The list contains 1.6 million molecular formulae. Since this collection was calculated *in silico*, the exact isotopic mass and natural isotopic abundance of the molecule (M) is known to many decimal places. Using this database we determined the subset of molecular formulae that contained at least one hydrogen atom and, removing this atom, generated a list of M-H fragments; roughly doubling the number of entries. Here it is interesting to note that in the removal of a hydrogen atom we have in effect generated a list of possible fragments where a single bond has been broken. Consider the case where a molecule is fragmented and a C-C bond is broken. Each fragment thus produced can be considered to be an M'-H species where the M' corresponds to the fragment if it were to have had a hydrogen attached/removed rather than being part of a larger structure. Therefore one can consider the list of M-H species to be all possible *fragments* of the same set of atoms (CHONSP). It is trivial to consider adducts of M; M+H, M+Na, M+K and the masses of these are calculated on demand. A software package – *SurfaceSpectra Identity* – that allows the user to search the 3.2 million molecular formulae and adducts for a spectral feature in the range 1-500 u, in the domain of CHONSP, and view the isotope pattern of the species is freely available from the SurfaceSpectra website[6]. In addition this software will display the isotope pattern of any other molecular formula, although this aspect does not determine whether the molecule is fully valent.

Data produced by electrospray ionisation (ESI), commonly used in LC-MS experiments where MS/MS is employed, have been shown to produce a large proportion of even electron (EE) ions in comparison to odd electron (OE) ions or radical cations[7]. A high EE/OE ratio compares well with the fragmentation pathway in SIMS data. This opens up the possibility of using LC-MS(-MS)^n libraries to assist in the interpretation of SIMS data. One such database is MassBank[8]. This website allows the user to search for a list of spectral features (peaks), given a mass tolerance, and returns a list of compounds that contain those peaks in any MS, MS^2, MS^3 or MS^4 data set, where available. The user can then select any of these 'hits' to extract the spectrum of the molecule, in as much detail as was originally uploaded by the contributor, and compare this to their SIMS spectra. There is unlikely to be a correlation between the relative intensity of the spectral features in the SIMS data and the LC-MS data, however the peak positions can give an insight into the likelihood of the SIMS data containing the molecule in question.

The MassBank organisation have also provided a web services interface to their collection. This allows for an automated search to be performed without the requirement for the user to

navigate the website. Software providers can therefore issue commands directly to the MassBank database and perform a series of commands without user intervention. Other databases of MALDI and LC-MS data include LipidMaps[9], Metlin[10] and Metfrag[11]; however these do not currently support a webservices interface.

Prediction

Given the (common) name of a molecule, or indeed a trade name, one can use the Royal Society of Chemistry's ChemSpider website[3] to determine the structure of the molecule together with a collection of its properties. For an increasing number of database entries there is also the possibility of a mass spectrum being available. Using the standard web interface, such spectra are presented in an interactive format and the underlying data can be downloaded for perusal off-line.

Given the molecular structure in a chemically meaningful format (SMILES[12], InChI[13] etc. rather than a 'picture' of the molecule, for example a jpeg) one can use one of the freely available open source chemical software libraries[14] to manipulate the structure. Now we are able to sequentially break each bond in the molecule and determine the molecular formula, and hence the isotopic distribution, of the resulting fragments. While these do not immediately represent the SIMS spectrum of the molecule, one can use this information as a starting point to suggest potential peak positions. In addition, using these open source packages one can perform rearrangements on the fragment structures, such as the loss of neutral species or other changes to bonding and structure that may be expected in the sputtering and subsequent relaxation processes, to produce other potentially useful peak positions. Of course, given that this is data calculated from knowledge of the elements involved, one can determine the isotopic distribution of the fragments with absolute accuracy.

ChemSpider presents a well developed webservices interface allowing a computer programmer to interact with the underlying database and to perform all calculations transparently on behalf of the user. The NIST Webbook[15] is another resource that allows the user to interact via a web page and contains some 15,000 mass spectra, albeit generated by electron impact ionisation.

Matching

Vector angle

Once one has identified a potential SIMS spectral standard that may be present as a component of the unknown spectrum, one requires a method of determining the likelihood of its presence. One method is to perform a spectral match using a vector angle approach. This technique, first developed over 30 years ago,[16] considers a spectrum as an N-dimensional vector, where N is the number of mass channels and the distance along each of these orthogonal directions is simply the intensity of the spectrum at that mass. If the standard and the unknown represent the same molecule, then within experimental error the vectors will be similar and the angle between these will be small (the cosine of the angle approaching unity); see Equation 1. Should the spectra represent different molecules then the differing intensities

at each mass will cause the vectors to diverge. Therefore one can consider the standard spectrum that presents the smallest angle (or largest cosine since one is often trying to maximise a match) to be the most appropriate match to the unknown.

$$\text{Cosine of vector angle} = \frac{\left(\sum LU\right)^2}{\sum L^2 \sum U^2}$$

Equation 1. Vector angle calculated from the vectors of the library spectrum L and the unknown spectrum U.

Of course, if there is no true match then the resulting outputs will all give a vector angle some distance from unity. Stein and Scott[17] considered a number of scaling parameters to weight the spectra and found that taking the square root of the spectral intensity and then scaling the result as a function of the cube of the mass, gave the best performance.

We selected 117 positive ion spectra of polymer additives from the *SurfaceSpectra Static SIMS Library*. These included 20 spectra of Irganox 1010 together with other Irganox molecules of similar molecular structure. Irganox 1010 was chosen by the curators of the SIMS Library as a test sample and as such was analysed by all contributors to the database. The 20 spectra span Ar^+, Ga^+, Xe^+, Cs^+ and C_{60}^+ primary ion sources and reflectron, TRIFT, Poschenrieder and quadrupole analyser types. The samples were prepared *in situ* and analysed in different laboratories around the world. Three of the samples were prepared on a silver substrate and as such produced cationised spectra.

One of these non-cationised Irganox 1010 spectra was selected to act as an 'unknown'. This unknown spectrum was then compared to each of the polymer additive spectra using the vector angle approach of Stein and Scott and the results are shown in Figure 1. The *y*-axis is a measure of the similarity of the spectra to the Irganox 1010 'unknown', where a value of 1 is considered a perfect fit, and the *x*-axis is the spectrum number in order of decreasing similarity value.

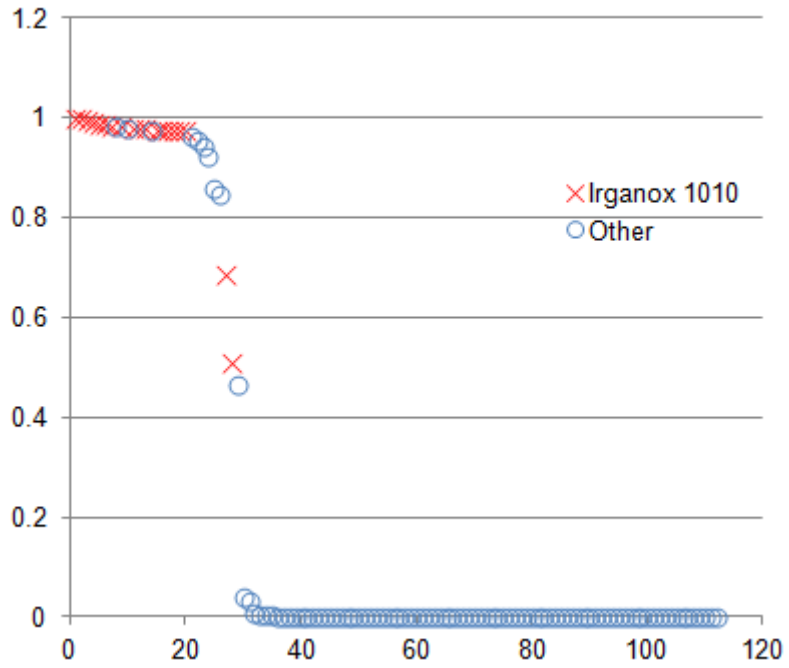

Figure 1 Result of vector angle match between various polymer additive spectra including 19 Irganox 1010 and a single Irganox 1010 spectrum acting as an 'unknown'.

It can be seen from the plot that most of the Irganox 1010 spectra indicated a very good fit with a vector angle value of greater than 90%. Two of the Irganox 1010 spectra exhibited a less good fit and, on investigation, it was found that these were cationised spectra. There are some other non-Irganox 1010 spectra that also exhibited a very good match and on analysis these are other Irganox type molecules that have the same structural formula, but in a different arrangement to the 1010 version.

This limited case suggests that the primary ion, whether cluster or atomic, and the mass analyser type may have little effect on the performance of this vector angle approach to spectral matching, giving us some confidence in using libraries of data from other laboratories. Further study including a variety of sample and material types and other ion source / analyser combinations is required to determine the stability of such an approach and where instability is found will be beneficial in understanding the spectral response in those cases.

The SurfaceSpectra Static SIMS Library search method

The SurfaceSpectra Static SIMS Library utilises a simple method of spectral matching. Each spectrum in the Library is processed as follows. Firstly the data are bin-summed to unit mass; intensity in channels from (n - 0.2 u) to (n + 0.8 u) is assigned to unit mass n. These peaks are then assessed against a threshold so that those with intensity below 100 counts/u are ignored; removing the potential for very weak peaks to confuse the search process. The spectrum is then broken down into 50 u segments and the ten most intense peaks in each segment are

recorded. It is this reduced set of peaks that the user matches against, with the number of ‘hits’ being presented to the user in descending order.

The reasoning behind this approach is twofold; to accommodate differences in mass resolution across the Library spectra and the user’s data and also to account for the wide range in transmission function present in SIMS data. A simple linear threshold of the data will be biased toward low mass, whereas diagnostically, it is the higher mass fragments that are most useful. These processes closely replicate those which a human interpreter would employ when first comparing an unknown spectrum with standard spectra. Differences in mass resolution and small errors in mass calibration will not affect the result.

Metadata

Informatics relies upon a store of well documented information that can be mined to produce correlations between unknown data sets and previously identified or standard data. Currently there is a lack of such well-documented information in the SIMS community.

In order to communicate efficiently with collaborating groups and potential users of our data in the future, it is essential to record information pertaining to the nature of the sample and experimental conditions in a common manner. This information is termed ‘metadata’; essentially, data about data. Here it is helpful to follow the lead of the biology community and use the concept of controlled vocabularies or ontologies, to define our information. An ontology, in this context, can be thought of as a tree structure of knowledge where each branch produces a more refined description of a piece of information.

Each term in an ontology has a formal description and is tagged using a unique formal identifier termed a Universal Resource Indicator (URI);[18] similar in effect to a URL used in defining websites. Indeed URLs are actually subsets of the URI framework. The use of a unique, text based identifier for a physical concept means that automated machine learning approaches can be used to search across data stores from different disciplines to determine related information.

For example, the imzML data transfer file format, suggested for imaging mass spectrometry[19], uses as its description format a modification of the mzML ontology developed by the Proteomics Standards Initiative of the Human Proteome Organisation (HUPO-PSI)[20]. In another example, the Royal Society of Chemistry’s Chemical Methods Ontology (CMO)[21] defines ‘secondary ion mass spectrometry imaging’ as follows:

```
[Term]
id: CMO:0000055
name: secondary ion mass spectrometry imaging
namespace: http://www.rsc.org/ontologies/CMO_OWL.owl
def: "The collection of spatially resolved mass spectra of a sample during
microscopy where the sample is bombarded with a stream of primary mass-
selected particles and the secondary ions ejected from the sample are
detected. The spectra are used to visualise the spatial distribution of
compounds by their molecular masses." [rsc:hjb]
synonym: "secondary ion mass spectroscopy imaging" EXACT []
synonym: "SIMS imaging" EXACT []
is_a: CMO:0000053 ! imaging mass spectrometry
```

This can be placed into a graphical description of the SIMS technique taken from the RSC's CMO ontology as depicted in Figure 2.

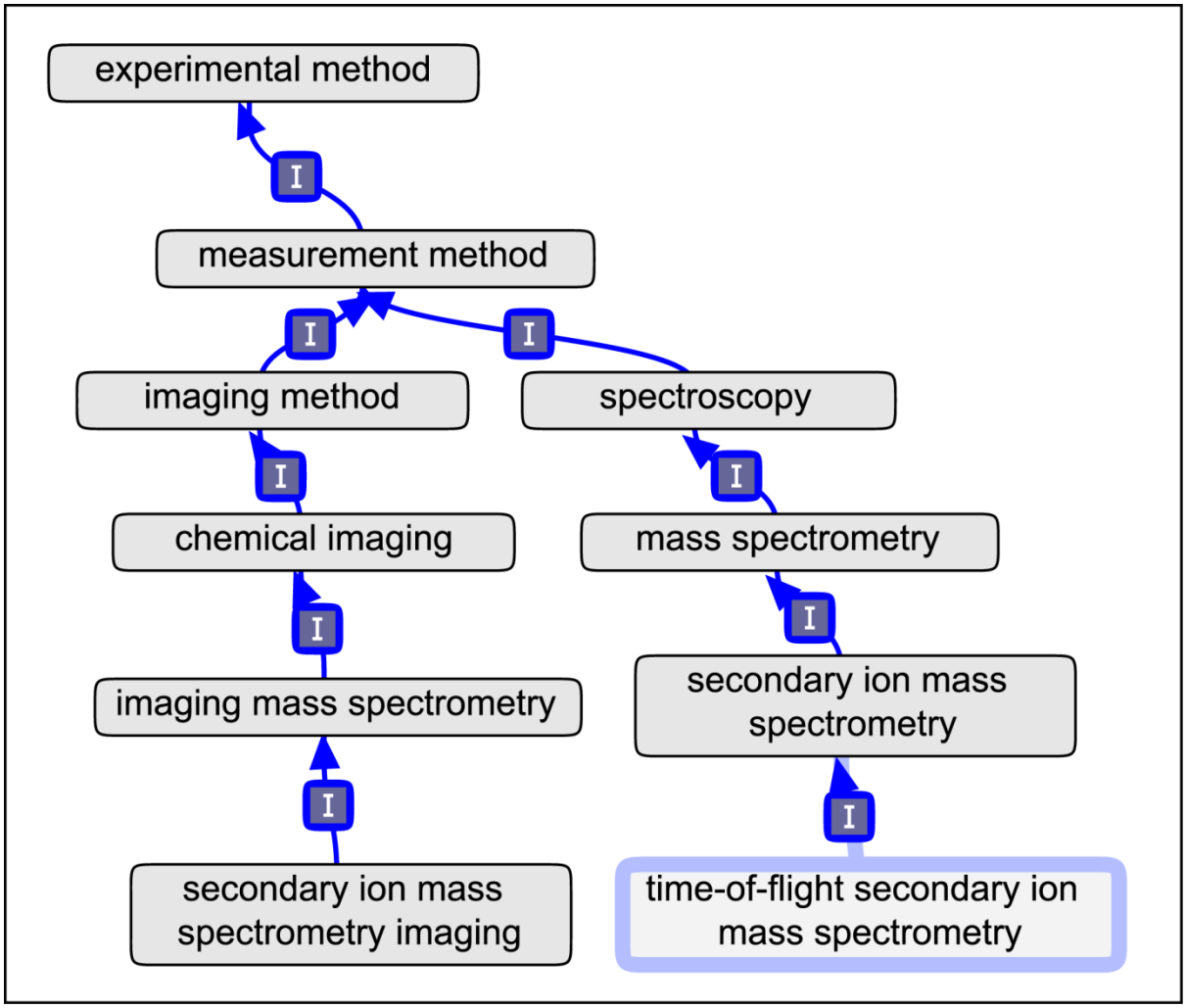

Figure 2 Graphical description of the hierarchy of SIMS in the RSC's CMO ontology. The 'I' annotation implies the concept of 'is a', for example imaging mass spectrometry 'is a' chemical imaging technique.

In order to facilitate the use of ontology terms by analysts, a Computer Science team from Manchester, in collaboration with the Heidelberg Institute for Theoretical Studies, developed a method of embedding ontology terms into spreadsheets.[22] The resulting software – *RightField* – allows an administrator to link a spreadsheet to one or more ontologies and constrain cells in the spreadsheet to offer selections from a subset of the ontology terms. This spreadsheet template can then be used by the analysts to record their experimental conditions in a formalised, but familiar manner.

Summary

We have shown that the area of informatics relating to the SIMS technique is broad and can benefit from connections with other mass spectrometries and computational approaches. The improving nature of these fields and gradual uptake of standardisation in experimental

annotation, suggests an expanding resource that SIMS practitioners will benefit from in the future. The field of SIMSInformatics may herald a step change in our approach to data interpretation.

Acknowledgements

The authors gratefully acknowledge the financial support of the Engineering and Physical Sciences Research Council, EPSRC, UK under grants EP/C008251 and EP/G045623/1. In addition we thank Tobias Kind and Oliver Fiehn for making their data openly available.

References

[1] I. S. Gilmore, M. P. Seah, F. M. Green, *Surface and Interface Analysis* **2005**, 37, 651. DOI: 10.1002/sia.2061

[2] *The SurfaceSpectra Static SIMS Library*, version 4 (Eds: J. C. Vickerman, D. Briggs and A. Henderson), SurfaceSpectra: Manchester, 2006. http://surfacespectra.com/simslibrary/ (retrieved March 2012)

[3] ChemSpider. http://www.chemspider.com/About.aspx (retrieved March 2012)

[4] CAS Registry. http://www.cas.org/expertise/cascontent/ataglance/ (retrieved March 2012)

[5] T. Kind, O. Fiehn, *BMC Bioinformatics* **2006**, 7, 234. DOI: 10.1186/1471-2105-7-234 http://fiehnlab.ucdavis.edu/projects/identification/ (retrieved March 2012)

[6] *SurfaceSpectra Identity*. http://surfacespectra.com/identity/ (retrieved March 2012)

[7] E. M. Thurman, I. Ferrer, O. J. Pozo, J. V. Sancho, F. Hernandez, *Rapid Communications in Mass Spectrometry* **2007**, 21, 3855. DOI: 10.1002/rcm.3271

[8] H. Hisayuki, A. Masanori, K. Shigehiko, N. Yoshito, I. Tasuku, S. Kazuhiro, O. Yuya, T. Kenichi, T. Satoshi, A. Ken, O. Yoshiya, K. Yuji, K. Miyako, T. Takayuki, M. Fumio, S. Yuji, H. Masami Yokota, N. Hiroki, I. Kazutaka, A. Naoshige, M. Takashi, T. Hiroki, A. Takeshi, S. Nozomu, S. Hideyuki, S. Daisuke, N. Steffen, I. Takashi, T. Ken, F. Kimito, M. Fumito, S. Tomoyoshi, T. Ryo, S. Kazuki, N. Takaaki, *Journal of Mass Spectrometry* **2010**, 45, 703. DOI: 10.1002/jms.1777 http://www.massbank.jp/ (retrieved March 2012)

[9] E. Fahy, M. Sud, D. Cotter, S. Subramaniam, *Nucleic Acids Research* **2007**, 35, W606. DOI: 10.1093/nar/gkm324 http://www.lipidmaps.org/ (retrieved March 2012)

[10] C. A. Smith, G. O. Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, G. Siuzdak, *Therapeutic Drug Monitoring* **2005**, 27, 747. http://metlin.scripps.edu/ (retrieved March 2012)

[11] S. Wolf, S. Schmidt, M. Muller-Hannemann, S. Neumann, *BMC Bioinformatics* **2010**, 11, 148. DOI: 10.1186/1471-2105-11-148 http://msbi.ipb-halle.de/MetFrag/ (retrieved March 2012)

[12] D. Weininger, *Journal of Chemical Information and Computer Sciences* **1988**, 28, 31. DOI: 10.1021/ci00057a005

[13] The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/inchi/ (retrieved March 2012)

[14] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, E. L. Willighagen, *Journal of Chemical Information and Modeling* **2006**, 46, 991. DOI: 10.1021/ci050400b http://blueobelisk.org (retrieved March 2012)

[15] NIST Mass Spec Data Center, S.E. Stein, director, "Mass Spectra" in NIST Chemistry WebBook, NIST Standard Reference Database Number 69, (Eds: P.J. Linstrom and W.G. Mallard), National Institute of Standards and Technology, Gaithersburg MD, 20899, http://webbook.nist.gov, (retrieved March 2012)

[16] Sokolow, S.; Karnofsky, J.; Gustafson, P. The Finnigan Library Search Program; Finnigan Application Report 2; Finnigan Corp.; San Jose, CA, March **1978**.

[17] S. E. Stein, D. R. Scott, *Journal of the American Society for Mass Spectrometry* **1994**, 5, 859. DOI: 10.1016/1044-0305(94)87009-8

[18] W3C, Naming and Addressing. http://www.w3.org/Addressing/ (retrieved March 2012)

[19] A. Römpp, T. Schramm, A. Hester, I. Klinkert, J.-P. Both, R. M. A. Heeren, M. Stöckli, B. Spengler, in *Data Mining in Proteomics, Methods in Molecular Biology*, Vol. 696 (Eds: M. Hamacher, M. Eisenacher, C. Stephan), Humana Press, **2011**, pp. 205-224. DOI: 10.1007/978-1-60761-987-1_12 http://www.maldi-msi.org/index.php?option=com_content&view=article&id=188&Itemid=63 (retrieved March 2012)

[20] mzML 1.1.0 Specification. http://psidev.info/index.php?q=node/257 (retrieved March 2012)

[21] Chemical Methods Ontology (CMO). http://www.rsc.org/ontologies/CMO/ (retrieved March 2012)

[22] K. Wolstencroft, S. Owen, M. Horridge, O. Krebs, W. Mueller, J. L. Snoep, F. du Preez, C. Goble, *Bioinformatics* **2011**, 27, 2021. DOI: 10.1093/bioinformatics/btr312 http://www.sysmo-db.org/rightfield (retrieved March 2012)