A scalable, scientifically validated workflow for biomarker identification and predictive toxicogenomics

Jens Hoefkens*, Juergen Cox, Marc Flesch, Tobe Freeman, Peter Haberl, Ruediger Heidenblut, Lynn Jablonski, Tom Jung, David Killen, Melanie Markmann, Kurt Zingler and Jim Samuelsson*
Genedata Inc. 1601 Trapelo Road, Suite 350 Waltham, MA 02451, USA. *Correspondence should be addressed to: Jens.Hoefkens@genedata.com

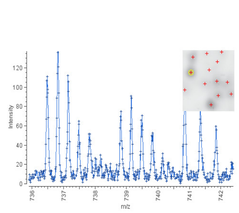
Advances in mass spectrometry (MS) now provide the possibility for sensitive high throughput biomolecule profiling of complex compound mixtures. Proteins and metabolites are of crucial importance to cellular processes and offer great potential as disease markers and as signatures of treatment side effects. Many proteins catalyze biological processes, and are therefore potential drug targets. However, a significant challenge lies in scaling up data processing and analysis methods for large and complex data sets.

We describe a highly automated workflow for high throughput MS data refinement and analysis. Developed in close collaboration with industrial partners, the workflow is implemented within an integrated bioinformatics platform and can be applied to proteomic as well as metabolomic data, and with all major technology configurations. We illustrate this workflow with an example from proteomics, using data acquired by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS).

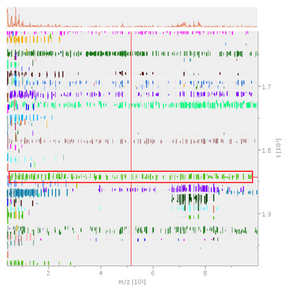
Approach

- Samples extracted from C. Elegans were spiked with known concentrations of bovine serum albumin (BSA).
- An automated workflow for refinement and analysis of LC-MS/MS data is demonstrated.
- In a blind study protein identities and concentration levels were correctly recovered.

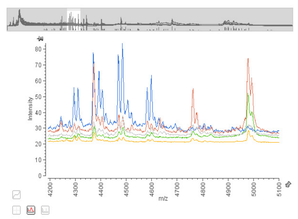
A Metabolomic and proteomic MS configurations supported by the software platform



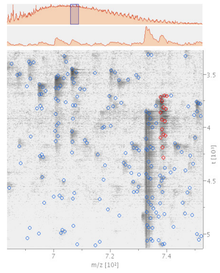
2D Gel and MS



GC-MS



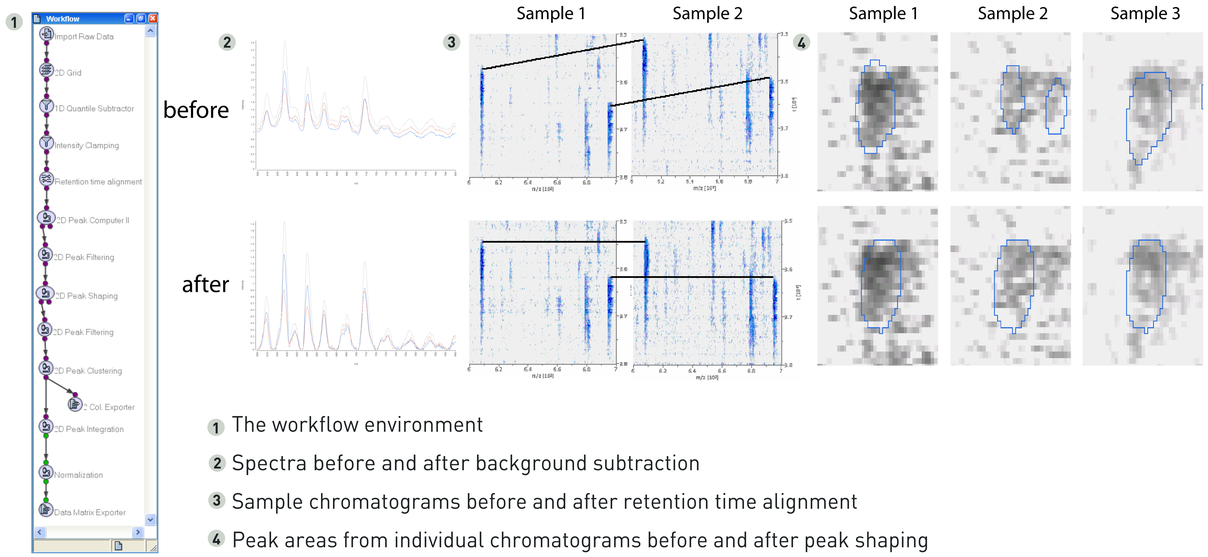
Individual spectra, e.g. SELDI data



LC-MS/MS

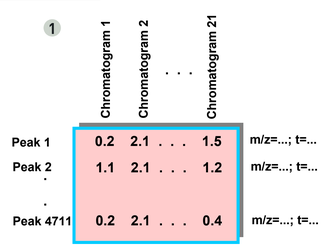
- Genedata's Expressionist software system for analysis of MS data supports all major MS technology configurations relevant for proteomics and metabolomics.
- Panels B-D focus on processing and analysis of LC-MS/MS data.

B Data refinement workflow applied to LC-MS/MS data

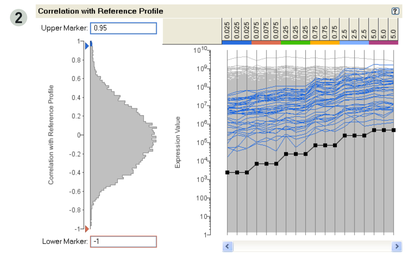


1. The workflow environment
2. Spectra before and after background subtraction
3. Sample chromatograms before and after retention time alignment
4. Peak areas from individual chromatograms before and after peak shaping

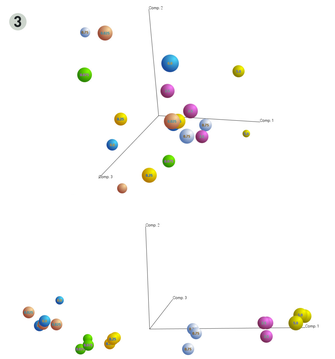
C Biomarker peak detection



Importing the data matrix for multivariate analysis. Refined data arranged into a matrix of extracted peaks (rows) versus sample chromatograms (columns).
Each matrix element corresponds to an individual peak volume (denoted by Expression Value in fig. 2).

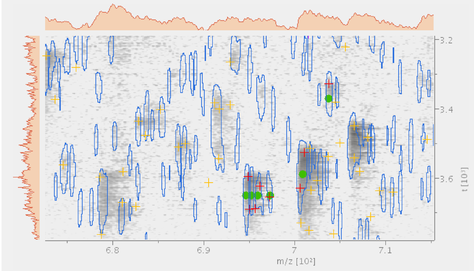


Each row in the data matrix forms a profile of expression values across chromatograms (gray lines). Retrieval of profiles (blue lines) showing strong correlation with the BSA spike concentration profile (black line). Profiles with a correlation coefficient above a user-defined threshold are defined as biomarkers and used for protein identification and concentration prediction.

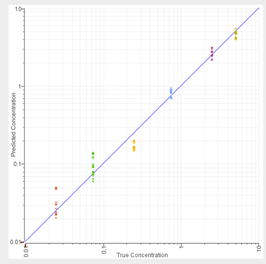


Principal Component Analysis (PCA) based on all peaks (upper) and based on the selected biomarkers only (lower).

D Validation of protein concentration and identification of protein



Retention time - m/z plane. Blue shapes are the boundaries of the peaks that were extracted in part B. Green dots indicate the biomarker peaks that were detected in part C. Crosses indicate the locations of precursor ions for which fragment spectra have been acquired.



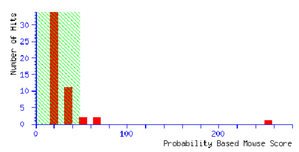
Given the ten most strongly correlated biomarkers a polynomial regression model predicts protein concentrations in cross validation. The full concentration range can be predicted accurately.

1. P02769-00-00-00 **Mass:** 69248 **Score:** 254 **Queries matched:** 5
(ALBU_BOVIN) Splice isoform Displayed; Variant Displayed; Conflict Displayed; from P02769
Check to include this hit in error tolerant search

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
3	582.77	1163.53	1162.62	0.90	0	49	0.04	1	K.LVNELTEFAK.T
8	653.90	1305.78	1304.71	1.07	0	33	1.7	1	K.HLVDEPQNLIK.Q
10	700.65	1399.28	1398.69	0.59	0	60	0.0051	1	K.TVMENFVAFVDK.C
16	740.78	1479.54	1478.79	0.75	0	62	0.0021	1	K.LGEYGFQNALIVR.Y
21	784.73	1567.45	1566.74	0.71	0	52	0.024	1	K.DAFLGSFLYEYSR.R

Mascot protein identification search result

- Those precursor ions which are detected as biomarkers are selected (red crosses in the upper left figure).
- Fragment spectra corresponding to biomarkers are automatically retrieved and submitted to a protein sequence database search engine (e.g. Mascot).
- Mascot search result yields clear protein identification.



Advantages

Versatile platform for MS-based proteomics and metabolomics

Developed in close collaboration with industrial partners

Identification of biomolecules and relative quantification

Genedata provides a genuine discovery tool for MS technologies

References

Scott D. Patterson: Data analysis – the Achilles heel of proteomics, Nature Biotech. 21, 221-222 (2003)

Hirai et al.: Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana, PNAS 101, 10205-10210 (2004)

Levander et al.: Automated methods for improved protein identification by peptide mass fingerprinting, Proteomics 4, 2594-2601 (2004)

Contact us at our website at www.genedata.com