# MINI-SYMPOSIUM ON STATISTICS AND PROTEOMICS

## Organizing committee

International Society for Clinical Biostatistics and Belgian Proteomics Association

## Objective

To introduce statisticians and sympathizers to the fascinating world of mass spectrometry-based proteomics.

## Program

| | |
|---|---|
| 8h00 | Registration |
| 8h50 | Short welcome |
| 9h00 | Tutorial I: Introduction to the technology and the database search engines |
| | Geert Baggermans, Dirk Valkenborg – The use of mass spectrometry in the context of proteomics |
| | Dirk Valkenborg, Kurt Boonen – Tandem MS peptide identification, search engines and confidence |
| 10h30 | Coffee Break |
| 11h00 | Tutorial II: Introduction to the quantification of proteins |
| | Lennart Martens – Approaches to MS-based metrics for protein quantification |
| | TBC, Lieven Clement – On the structure of the data and protein inference |
| 12h30 | Lunch |
| 13h30 | Selected topics I: The future of methodology |
| | Bart Mertens – Pitfalls and opportunities in statistical learning applications for mass spectrometry-based clinical proteomics |
| | Ralf Gabriels, Sven de Groeve – MS$^2$PIP: Predicting peptide spectrum peak intensities to improve proteomics identification |
| | Maarten Dhaesens – DIA: next level proteomics |
| | Gerben Menschaert - Proteogenomics data integration |
| 15h00 | Coffee Break |
| 15h30 | Selected topics II: Moving towards the protein level |
| | Wim Vranken – Towards a statistical view of protein behaviour |
| | Jürgen Claesen – The use of the isotope distribution in statistical models for MS data: an example from structural proteomics |
| | Nico Verbeeck – Mass spectrometry imaging and machine learning, a fruitful marriage |
| | Laurent Gatto – Probabilistic modelling of protein sub-cellular localization |
| 17h00 | Closing of the symposium |

# Tutorial Abstract:

Identification (Geert Baggerman, Dirk Valkenborg, Kurt Boonen)

Proteins were discovered earlier than DNA, yet in the biotechnology community it is perceived that the genomics or transcriptomics field is currently leading the scientific breakthroughs. In this tutorial you will learn about the importance of the field of proteomics, their closeness to the phenotype and the breakthroughs that have brought this discipline into the high-throughput technology field. Much attention is given to the identification of proteins – a first step in analysis the high-throughput data flood.

Quantification (Lennart Marten, Lieven Clement):

There are many ways in which proteins can be quantified, with different approaches being used in the design and execution of the experiment.

These methods will first be situated according to their underlying concepts, and then each will be briefly described, along with its applications and current popularity. What all these methods share, however, is the use of the mass spectrometer as a proxy readout for the quantity of the protein in the original sample. This process comes with typical detector issues, which will be explained, and moreover also with some proteomics-specific caveats that will be highlighted as well.

# Selected Topics Abstracts

Bart Mertens:

Modern proteomics depends crucially on mass spectrometry to simultaneously investigate the expression of hundreds of proteins in clinical studies of tissue or blood samples. While it is a powerful technology which generates high-dimensional data, validity of statistical inference depends crucially on appropriate statistical design. In this talk we briefly discuss some common pitfalls in study design for mass spectrometry-based measurement. We identify the key objectives of statistical study design for mass-spectrometry proteomics.

Ralf Gabriels:

In mass spectrometry-based proteomics, peptides are identified by analyzing their fragmentation spectra. While the first dimension of a given peptide's fragmentation spectrum - the peak m/z values – are easily calculated, the spectrum's second dimension - the peak intensities - follows a more complex pattern. As a result, traditional proteomics search engines mainly focus on m/z values when identifying peptide spectra. In this talk, I will show how we developed MS$^2$PIP, a machine learning application that accurately predicts peptide spectrum peak intensities, and how this can improve peptide spectrum identifications.

Maarten Dhaesens:

Data-independent acquisition (DIA) currently generates the most comprehensive mass spectrometry (MS) data. In this strategy, the instrument continuously cycles through predefined mass over charge ratio (m/z) windows to regularly measure the intensity of peptide fragments throughout the liquid chromatography (LC)

gradient. This is considerably richer than conventional data-dependent acquisition (DDA), where only the most abundant peptides are stochastically fragmented. However, the DIA data is chimeric to such an extent that the extraction of relevant information has shown to be very challenging. Thus, there is an increasing tendency to think ion-centric, as opposed to peptide-centric. This implies a manifold increase in data complexity. With each ion retaining its individuality, this calls for increasingly more extended statistical methods. But only little has been done so far…

Gerben Menschaert:

TBC

Wim Vranken:

TBC

Jürgen Claesen:

In high-resolution mass spectrometry (MS), a peptide appears in a mass spectrum as a series of locally correlated peaks, which exhibit a specific characteristic profile related to the isotope distribution of the peptide. The isotope distribution reflects the number and probabilities of occurrence of different isotopic variants of a molecule. The probabilities are reflected in a mass spectrum by the relative heights of the series of peaks related to the molecule; whilst the different masses result from the fact that there are different isotopes of chemical elements.

The use of the (expected) isotope peak-patterns can increase effectiveness of selecting the relevant information from mass spectra and subsequent statistical analysis of the data. In this presentation, I will give an example how information about the isotope distribution in the analysis of peptide-centric HDX-MS data can result in relevant protein-structure information.

Nico Verbeeck:

TBC

Laurent Gatto:

TBC