Poster G-12

Beating the Noise: New Statistical Methods for Detecting Signals in MALDI-TOF Spectra below Noise Level

Authors:
Tim OF Conrad (*Free University of Berlin, Department of Mathematics and Computer Science*)
Alexander Leichtle (*Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, University Hospital*)
Andre Hagehuelsmann (*Microsoft Research*)
Elmar Diederichs (*Free University of Berlin, Department of Mathematics and Computer Science*)
Sven Baumann (*Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, University Hospital*)
Joachim Thiery (*Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, University Hospital*)
Christof Schuette (*Free University of Berlin, Department of Mathematics and Computer Science*)

Short Abstract: A new analysis workflow for MALDI-TOF massspectrometry data has been developed capable of identifying and analyzing signals (peaks) even below noise level and accurately determination of their parameters.We believe that this will foster identification of new biomarkers having not been detectable by most algorithms currently available.

Long Abstract:

Background:
The computer-assisted detection of small molecules by mass spectrometry in biological samples provides a snapshot of thousands of peptides, protein fragments and proteins in biological samples. This new analytical technology has the potential to identify disease associated proteomic patterns in blood serum. However, the presently available bioinformatic tools are not sensitive enough to identify clinically important low abundant proteins as hormones or tumor markers with only low blood concentrations.

Aim:
Find, analyze and compare serum proteomic patterns in groups of human subjects having different properties such as disease status or epidemiological parameters (e.g. gender, age) with a new workflow to enhance sensitivity and specificity.

Problems:
Mass data acquired from high-throughput platforms frequently are blurred and contain high- and low-frequency (baseline) noise. This complicates the reliable identification of peaks in general and very small peaks below noise level in particular. It is due to the fact that it is not possible to distinguish noise from signals if these two components fully overlay. However, this statement is only valid for single or few spectra. If the algorithm has access to a large number of spectra (e.g. N > 1000), new possibilities arise, one of such being statistical

methods.

Approach:
Apply signal preprocessing steps followed by statistical analyses of the blurred data and the region below the typical noise threshold to identify signals usually hidden below this ``barrier''. The steps applied are (briefly summarized):

(1) Baseline Correction: a morphological Top Hat filtering is applied to remove the baseline from the raw data acquired from the mass-spectrometry machine.

(2) Signal Smoothing is achieved by application of level-dependent wavelet shrinkage techniques.

(3) For spectra normalization a modified total ion current (TIC) approach is utilized.

(4) To identify peaks, properties of the Top Hat Filter are exploited, that is, the area of intersects of the signal with the x-axis are treated as (presumably convoluted) candidate peaks.

(5) Peak Deconvolution is conducted through statistical test based application of Kernel Density Estimation techniques and subsequent analysis of the components found.

(6) Determination of peak properties such as area or quality are determined through the usage of sophisticated mathematical concepts, e.g. Fourier integral calculation or geometrical hashing for shape assessment, respectively.

(7) Peak assignment across spectra in the same spectra group (such as diseased, healthy,low-concentration, or high concentration) is performed using the Chinese Restaurant Process (CRP) to cluster similar peaks appearing across spectra.

Deliverables:
A new analysis workflow has been developed capable of identifying and analyzing peaks even below noise level and accurately determination of their parameters. These peaks can be used in subsequent steps to build peak patterns for proteomic pattern analysis. We believe that this will foster identification of new biomarkers having not been detectable by most algorithms currently available. First steps towards distributed computing approaches have been incorporated in the design.

Application Experiment:
To obtain a first ``proof-of-principle'' and to test the overall performance of our workflow we spiked a small subset of known human serum samples with a protein standard mix (concentrations: 127nMol/L, 0.79nMol/L, 0.32nMol/L, 3.17pMol/L, 0.079pMol/L) leading to five different concentration groups. From these mixtures spectra were obtained by an Autoflex Linear MALDI-TOF Mass Spectrometer (Bruker Daltonics, Germany) and processed by our pipeline. For each of the five resulting concentration groups we evaluated whether and where the spiked substances could be found.

Summary of Application Results:

First, the algorithms successfully detect peaks even for very small concentrations at pMol/L level, second, the methods are able to detect peak centers very accurately even in very small concentrations where other platforms fail to detect these proteins. With our algorithms it was possible to detect the hormones Angiotensin, ACTH clip 18-39, Substance P and the cell protein Ubiquitin at very low and clinically relevant concentrations.