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

T. OF Conrad[1], A. Leichtle[2], S. Baumann[2], E. Diederichs[1], U. Ceglarek[2], G.M. Fiedler[2], A. Hagehülsmann[3], J. Thiery[2], C. Schütte[1]

1 Freie Universität Berlin
2 Universität Leipzig, Universitätsklinikum Leipzig, Anstalt öffentlichen Rechts, IFM
3 Microsoft Research

Introduction

Background: Mass spectrometry based screening methods have been recently introduced into clinical proteomics. This boosts the development of a new approach for early disease detection: proteomic pattern analysis.

Aim: Find, analyze and compare proteomic patterns in groups of patients having different properties such as disease status or epidemiological parameters (e.g. sex, age) with a new pipeline to enhance **sensitivity and specificity**.

Problems: Data acquired from high-throughput platforms frequently are blurred and noisy. This extremely complicates the reliable identification of peaks in general and very small peaks below noise-level in particular.

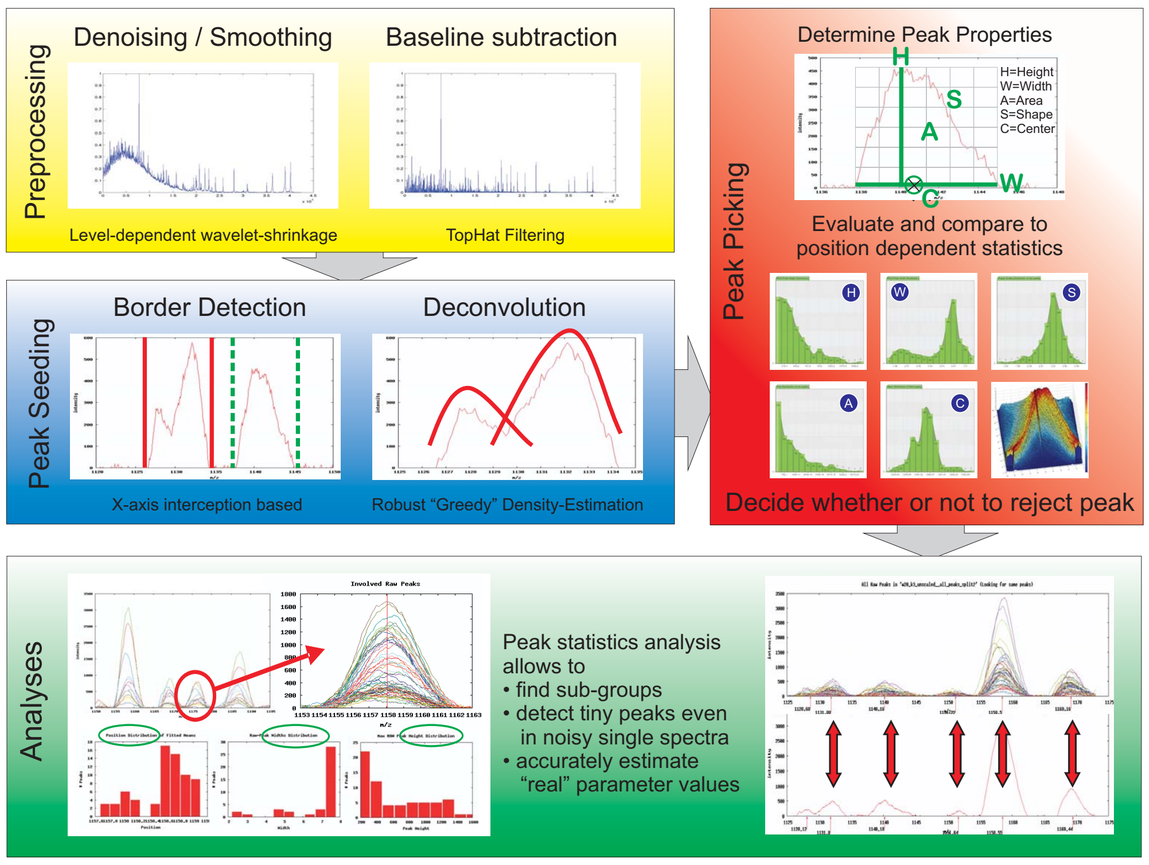
Approach: Apply sophisticated **signal preprocessing** steps followed by **statistical analyses** to purge the raw data and enable the detection of real signals while maintaining information for tracebacks.

The puzzle images illustrate the fundamental problem in proteomics driven biomarker discovery: To identify relevant traces of biomarkers within a vast abundance of peptides and proteins. The correctness of classification mainly depends on technical parameters (resolution, mass shift, spectrometer settings) as well as on preprocessing and classification quality. The puzzle images above differ in five characteristics - the importance of detection sensitivity and accuracy becomes intelligible.

Data

Samples: Serum samples of more than 700 apparently healthy blood donors from a previously published study served as basis for the investigation [2]. After centrifugation and aliquotation samples were stored at -80°C until analysis.
Preparation: After thawing on ice peptide and protein purification and fractionation was performed using a magnetic bead-based separation technique (ClinProt™, Bruker Daltonics) availing specific surface functionalities (MB-IMAC Cu, MB-WCX, MB-HIC 8) as predetermined in our related publication [3].
Chemicals & Consumables: Gradient grade acetonitrile (ACN), ethanol, acetone were obtained from J.T. Baker (Phillipsburg, USA); trifluoroacetic acid (TFA) was purchased from Sigma-Aldrich (Steinheim, Germany). Peptide preparations were done in 0.2 ml polypropylene tubes (8-tube strips) from Biozym (Hess. Oldendorf, Germany). The MALDI-TOF AnchorChip™ target was purchased from Bruker Daltonics (Leipzig, Germany). Peptide and protein calibration standards, a-cyano-4-hydroxycinnamic acid (HCCA) were purchased from Bruker Daltonics (Leipzig, Germany). Peptide admixture series in serum were perfomed using peptide calibration standard No. 206195 (Bruker, Leipzig, Germany). Standard concentrations in serum (16 samples/level) ranged from 0.008pmol/l to127nmol/l.
Hardware Configuration: ClinProt™ Robot & Autoflex Linear MALDI-TOF Mass Spectrometer (Bruker Daltonics, Germany)
Mass Spectrometry: Mass spectra were recorded by the flexControl™ 2.0 Software (Bruker Daltonics, Germany). The settings were applied as follows: Ion source 1: 20 kV; ion source 2, 18.50 kV; lens, 9.00 kV; pulsed ion extraction, 120 ns; nitrogen-pressure, 2500 mbar. Ionization was achieved by a nitrogen laser (λ=337 nm) operating at 50 Hz. For matrix suppression a high gating factor with signal suppression up to 500 Da was used. Mass spectra were detected in linear positive mode. Spectral data were combined with beforehand surveyed epidemiological and clinical metadata in a Microsoft SQL Server™ database to provide highly differentiated classification criteria.

Pipeline



Results

Improvement in Peak Identification...

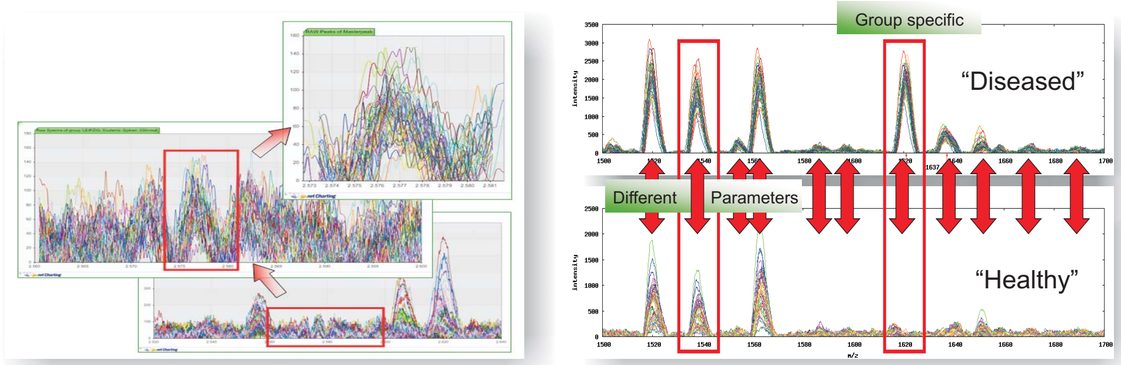
- Create statistical information about peak "clusters" located in one spot across spectra.
- Use this information to find sub-groups in overlapping peaks and peaks even below noise-level.

⇨ We were able to gain significant improvement in peak identification sensitivity compared to proprietary software (ClinProTools™ 2.0β, Bruker Daltonics, Leipzig, Germany). Independence of noise-levels allows us to find peaks in a peptide-mix in a concentration as low as 3.17pmol/l, compared to a minimum concentration of 0.32nmol/l needed by ClinProTools™2.0β.

Substance	Plat-form[3]	Theor. Center	Center found (Concentration of spiked peptide mix as stated below)				
			121.21nMol/L	0.76nMol/L	0.30nMol/L	3.03pMol/L	0.075pMol/L
Angiotensin II	P.NET	1047.20	1047.1	-[1]	-[1]	-[1]	-[1]
Angiotensin II	CPT	1047.20	1046.9	-[2]	-[2]	-[2]	-[2]
Angiotensin I	P.NET	1297.51	1297.6	1298.0	1299.3	1299.2	-[1]
Angiotensin I	CPT	1297.51	1297.2	-[2]	-[2]	-[2]	-[2]
Bombesin	P.NET	1620.88	1620.9	1618.1	1617.2	1617.2	-[1]
Bombesin	CPT	1620.88	1620.6	-[2]	-[2]	-[2]	-[2]
ACTH clip 18-39	P.NET	2466.73	2466.8	2465.8	2465.8	2466.2	-[1]
ACTH clip 18-39	CPT	2466.73	2466.2	-[2]	2466.1	2465.9	-[2]
Somatostatin 28	P.NET	3149.61	3149.5	-[1]	-[1]	-[1]	-[1]
Somatostatin 28	CPT	3149.61	3149.0	-[2]	-[2]	-[2]	-[2]
Insulin	P.NET	5734.56	5734.3	-[1]	-[1]	-[1]	-[1]
Insulin	CPT	5734.56	5734.2	-[2]	-[2]	-[2]	-[2]

Table 1. Results of the spiking experiments (see text for explanation). Note that no calibration has been performed for the Proteomics.NET platform.
[1]: No significant masterpeak in this range at this concentration found, [2]: No significant peaks in this range at this concentration found, [3]: CPT: Bruker ClinprotTools, P.NET: Our Proposed Platform: Proteomics.NET

... Fosters Identification of very fine-grained Differentiation Patterns



Conclusion

By adapting and sensitizing the pipeline parameters and methods, we found a **significant gain of sensitivity and specificity for the succeeding classification steps**, even compared to proprietary software. We are able to **identify single peaks allowing to significantly distinguish sub-groups** of patient spectra having different meta data (e.g. gender, age, or peptide admixture). Since clinical and epidemiological data frequently are highly categorizable, classification systems incorporating multiple-class affiliation will ameliorate the search for unique class-specific spectral features.

This work is supported by grants of the Sächsische Aufbaubank (SAB) and Microsoft Research.

[1]: Wagner M, Naik D, Pothen A: Protocols for disease classification from mass spectrometry data. Proteomics. 2003, 3(9):1692-8.
[2]: Kratzsch J, Fiedler GM, Leichtle A, Brugel M, Buchbinder S, Otto L, Sabri O,Matthes G, Thiery J: New reference intervals for thyrotropin and thyroid hormones based on National Academy of Clinical Biochemistry criteria and regular ultrasonography of the thyroid. Clin Chem. 2005, 51(8):1480-6.
[3]: Baumann S, Ceglarek U, Fiedler GM, Lembcke J, Leichtle A, Thiery J: Standardized approach to proteome profiling of human serum based on magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clin Chem. 2005, 51(6):973-80.