Fishing for biomarkers: analyzing mass spectrometry data with the new ClinProTools™ software

Ralf Ketterlinus[1], Sen-Yung Hsieh[2], Shih-Hua Teng[3], Helen Lee[2], and Wolfgang Pusch[1]
[1]Bruker Daltonik GmbH, Bremen, Germany, [2]Chang Gung Memorial Hospital, Tao-Yuan, and [3]Bruker Daltonics Ltd., Taipei, Taiwan

BioTechniques 38:S37-S40 (June 2005)

Recently, applications of mass spectrometry in the field of clinical proteomics have gained tremendous visibility in the scientific and clinical community. One major objective is the search for potential biomarkers in complex body fluids like serum, plasma, urine, saliva, or cerebral spinal fluid. For this purpose, efficient visualization of large data sets derived from patient cohorts is crucial to provide clinical experts an interactive impression of the data quality. Additionally, it is necessary to apply statistical analysis and pattern matching algorithms to attain validated signal patterns that may allow for later applications in sample classification. We introduce the new ClinProTools™ bioinformatics software, which performs all major steps of profiling, screening, and monitoring applications in clinical proteomics. ClinProTools is the data interpretation software of the mass spectrometry-based ClinProt™ solutions for biomarker analysis. ClinProTools performs data pretreatment, visualization, statistics, pattern determination, pattern evaluation, and classification of spectra. This article will focus on ClinProTool's powerful and intuitive visualization options for clinical proteomics applications.

INTRODUCTION

Mass spectrometry-based search for biomarker patterns is widely recognized as a valuable research tool for predictive medicine and pharmacological monitoring (1). Of particular interest is the identification of tumor markers for early detection and the diagnosis of cancer to improve the clinical prognosis of patients. Various disciplines may take advantage of this new technique in their clinical studies (i.e., oncology, urology, psychiatry, neurology, toxicology, pharmacology, and others). Additionally, this application may even be adopted beyond clinical applications (i.e., from food and seed production to pathogen and mycotoxin profiling).

The pursuit to recognize specific polypeptide biomarker patterns for certain diseases adds a new level to proteomics that demands sophisticated evaluation (2) and valuable statistical tools. Advances in sample preparation and instrumentation enhance the requirements for comprehensive computational methodologies to screen and evaluate data sets. Universal, intuitive, and flexible bioinformatics solutions are needed to satisfy current and upcoming analytical needs.

MATERIALS AND METHODS

Sample Collection and Preparation

Serum samples were collected from 13 patients who were admitted to our hospital because of pneumonia. Meanwhile, two groups of other patients with approximate age and gender matching were enrolled as controls, including 16 patients presented with fever, but without clinical or radiological evidence of pneumonia, and 13 obviously healthy individuals as disease controls and healthy controls, respectively. Serum samples were collected in acute phase of disease for the pneumonia patients, and during the episodes of fever for the disease control group. All serum samples were collected after at least 8 h fasting. Sera were snap-frozen and stored at -80°C in aliquots until use. All experiments were conducted in duplicate.

Prefractionation

Prefractionation has been performed using different surface functionalities of ClinProt™ microparticle beads (Bruker Daltonik, Leipzig, Germany) to generate a diversity of biomarker patterns for the profiling. We have used hydrophobic interaction (MB-HIC8), weak cation ion exchange (MB-WCX), and immobilized metal-affinity chromatography containing copper ions (IMAC-Cu). All preparations have been performed according to the manufacturer's instructions.

Robotics

The complete magnetic bead fractionation has been performed on an automated robotic platform (ClinProt Robot; Bruker Daltonik). This included all pipeting steps and matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) sample plate loading.

Mass Spectrometry

Mass spectrometry has been performed using the ClinProt system, equipped with an Ultraflex™ TOF/TOF instrument (Bruker Daltonik). Spectra have been collected automatically using the AutoXecute™ software (Bruker Daltonik; see Reference 3) for fuzzy-controlled adjustment of critical instrument settings to generate raw data of optimized quality.

Data Interpretation

The ClinProTools™ software (Bruker Daltonik) has been used for all data interpretation steps, which start with a raw data pretreatment, including normalization of a set of spectra derived from a patient cohort, internal signal alignment using prominent internal signal peaks, and a peak picking procedure. The whole data pretreatment has been completed using default settings and was performed automatically, without any user interaction. The pretreated data have been used for visualization and statistical analysis in ClinProTools. Peak statistics has been performed by means of a Welch's *t*-test.

RESULTS

The ClinProTools software was used for the data interpretation of MALDI-TOF spectra derived from serum samples of different patient groups. A subset of the patients has been affected by the target disease, pneumonia. We have used two different control groups, namely disease controls and healthy controls, which enabled us to distinguish general disease marker candidates from those that

are specific for the target disease. The samples have been prefractionated using magnetic beads with three different surface functionalities, namely MB-HIC8, MB-WCX, and MB-IMAC-Cu.

Visualization of Mass Spectrometry Data

ClinProTools offers a variety of viewer options for the analysis of clinical profiling data (Figure 1). A virtual gel view gives an overview representation of data sets derived from large sample cohorts. This viewer is the master navigation tool for a large data set. The virtual gel view can be displayed in different color scales. A representation of the signal intensities in a rainbow color scale allows for detection of faint differences between the classes due to the changing colors with increasing peak height (Figure 1).

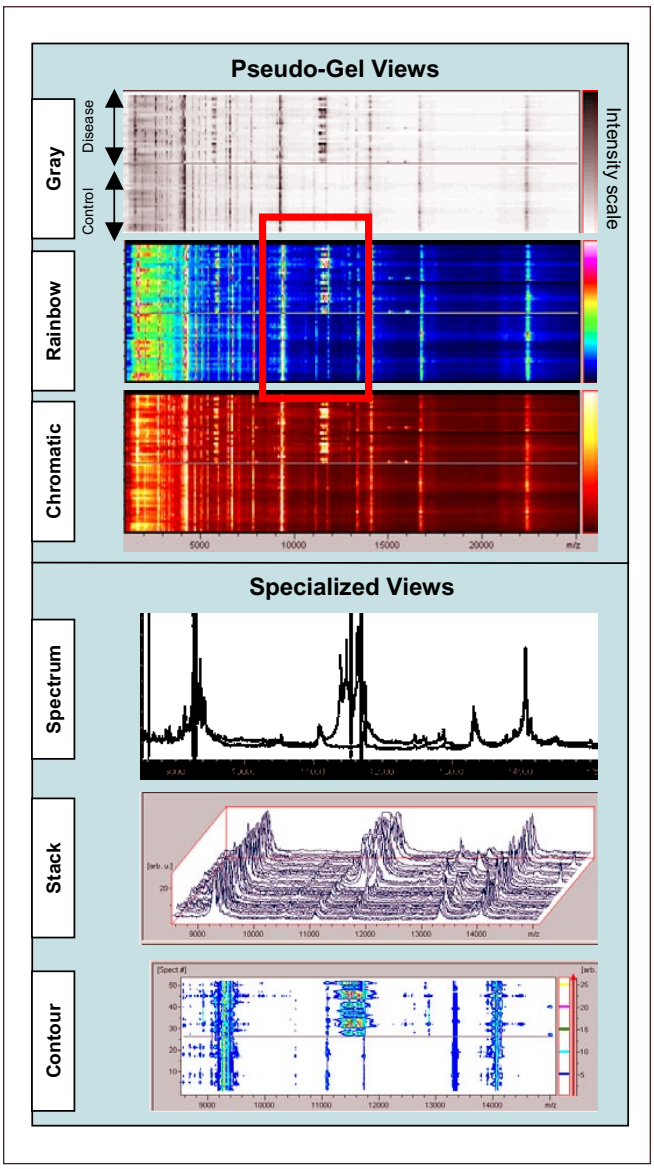


Figure 1. Different viewer options in the ClinProTools bioinformatics package. The pseudo-gel view is the master navigation tool in a data set. All individual spectra are shown in a density scale. Both groups (target disease and healthy control) contain 13 samples that have been prepared in duplicate. The specialized views were zoomed into the region of approximately 9000–14,000 Da with multiple differentially displayed signals. The corresponding mass region is labeled with a red frame on the left panel. The 3-dimensional stack view gives further hints concerning signal scattering in a sample class. In the contour view, different intensity thresholds can be defined by the user, and the data are represented similar to a topographical map with contour lines.

An alternative viewer shows a conventional spectrum display, which can be customized to use individual spectra or prototype spectra of the different sample classes. After peak selection, a "box & whiskers" plot graphically displays the intensity distribution of the relevant signals in the different sample classes.

Further options are offered by a stack plot viewer to get an impression of signal intensity scattering in a 3-dimensional view and in a contour plot viewer that allows defining multiple intensity threshold in the display.

Classical and Modern Approaches to Detect Pattern Differences: Peak Statistics and the Genetic Algorithm

Two independent statistical approaches assist the user in the detection of peak pattern differences of assigned spectra to the different classes, namely classical statistics of peak signal intensities and a modified genetic algorithm. Moreover, data export allows applying other statistical approaches like decision tree analysis (Salford Systems, San Diego, CA, USA) or in-house bioinformatics.

ClinProTools provides a list of peaks sorted according to the statistical significance to differentiate between both classes (Figure 2). On the basis of a Welch's *t*-test, a *P* value is calculated, which indicates the probability that the observed intensity differences of the individual peaks are not based on coincidence. These calculations are done independently for peak heights and peak areas. A list of all peaks sorted according to the statistical separation strength will be created as an output file. Figure 2 shows the results for two classes of the test serum profiles. For examination, the respective masses can easily be inspected in the spectrum view and in the virtual gel view (Figure 2).

User Interactivity, Validation, and Class Prediction

Peaks with high separation power in the Welch's *t*-test may be used to generate a biomarker pattern model. Peaks may also be added manually by the user. Classical peak statistics and the genetic algorithm may be used in combination. Once saved, pattern models may be validated in a cluster analysis with an independent test sample set.

If the user did not already establish an independent test data set of sufficient cohort size for a validation, cross validation (4) of the cluster analysis may be used to investigate the quality of specific signal patterns. For the validation of a model, independent test data sets representing the classes can be selected by the user. Validation determines the predictive capability of a model as a percentage of the correctly classified test data. Finally, new unknown samples may be classified in a class prediction, which can be carried out with a previously established pattern model.

ClinProTools accompanies the user through the whole process—from acquiring reasonable biomarker patterns, to the final step of classification of unknown samples. Comprehensive visualization tools, supplemented with sophisticated mathematical algorithms, deliver reproducible results that can instantly be validated and used in class prediction.

Interpretation of Data Derived from Pneumonia Patients

We have performed an initial study concerning visualization and statistical analysis of a data set containing 52 spectra derived from duplicate preparations of 13 disease patients and 13 healthy controls. The aim of this study was to test the visualization capabilities of the software for the identification of some biomarker candidates, which would be promising for a more detailed analysis in subsequent studies with a real life data set. It was explicitly not the

aim of this study to establish diagnostic biomarker patterns; a larger data set is needed for such a confirmation. Moreover, identification via TOF/TOF fragment analysis should be performed to provide biological relevance to the statistical analysis.

We have found several mass regions with differences between the two sample classes upon visual inspection (Figures 1 and 2). When using WCX beads for the prefractionation, one obvious region of multiple signals showed differential intensities between the two classes in the range of 11,500 Da. Zooming into this region revealed that it is composed of multiple differential signals (Figure 1). Upon detailed inspection of the complete data set, further differential signals have been found in the mass range from 1500 to 10,000 Da (data not shown). Subsequently, we have used the statistical features of ClinProTools to evaluate those peaks with seemingly high power to differentiate between the two classes by means of a Welch's *t*-test. The result of this test is a *P* value, which gives an estimation of the probability that the measured peak signal distributions can be observed by chance. Accordingly, the lower the *P* value, the better a respective peak signal is suited to be used to separate the two classes. The generally accepted limit of the *P* values to consider a result significant is defined as 0.05 (respectively, 0.01 for highly significant results). The output table of the statistical analysis in Figure 2 shows that several peaks in the data set resulted in *P* values indicating high importance. The combined availability of visualization and statistical analysis allows a direct feedback to the original data. Accordingly, the peaks labeled as significant to distinguish between the two classes have been directly controlled using the visualization features. This is shown in Figure 2 for the two peaks at 4662 and 11,525 Da. In both cases, the peaks of interest show different signal intensities between the two classes, while adjacent peaks have nearly identical intensities, thus serving as internal controls (see asterisks in Figure 2).

It is of special interest to distinguish between biomarker candidates for a general disease condition and those that are specific for the target disease. To demonstrate the respective software functionality, we have used three independent groups, namely healthy control, disease control, and target disease. Figure 3 shows the combined results after fractionation using the hydrophobic C8 magnetic beads. Statistical analysis gives a number of peaks, which can be used to separate healthy controls from both disease groups (labeled red in Figure 3). Some further signals are specific for the target disease, as they can be used to separate the respective samples of both control groups (labeled green in Figure 3). These different kinds of potential biomarkers can be especially well visualized in the mass region between 3100 and 3400 Da. Here we found biomarker candidates of both types.

DISCUSSION

The presented initial data interpretations of samples derived from pneumonia patients demonstrate that the combined visualization and statistical features of the ClinProTools software allows for an easy and efficient identification of biomarker candidates. Such candidates can later be used in

Index	AveMass	Ave1	Ave2	DAve	StdDev1	StdDev2	Conf1	Conf2	p-value	t-test
186	4661.69	75.83	32.69	43.15	29.53	23.99	28.14	58.15	5.382e-007	5.78
171	4234.96	568.73	267.88	300.87	173.34	209.75	193.59	408.15	8.77e-007	5.64
287	9290.59	562.62	291.29	271.34	204.02	142.38	173.04	369.63	1.42e-006	5.56
328	11524.9	-0.14	35.85	-35.99	3.02	34.74	-50.07	-21.92	1.81e-005	-5.26
269	8630.33	-0.31	0.77	-1.08	0.86	0.77	-1.53	-0.62	1.81e-005	-4.75
214	5769.97	-0.89	15.13	-16.02	3.15	15.85	-22.52	-9.52	2.63e-005	-5.06
66	2137.57	5.16	21.92	-16.77	8.73	16.09	-24.03	-9.50	3.6e-005	-4.67

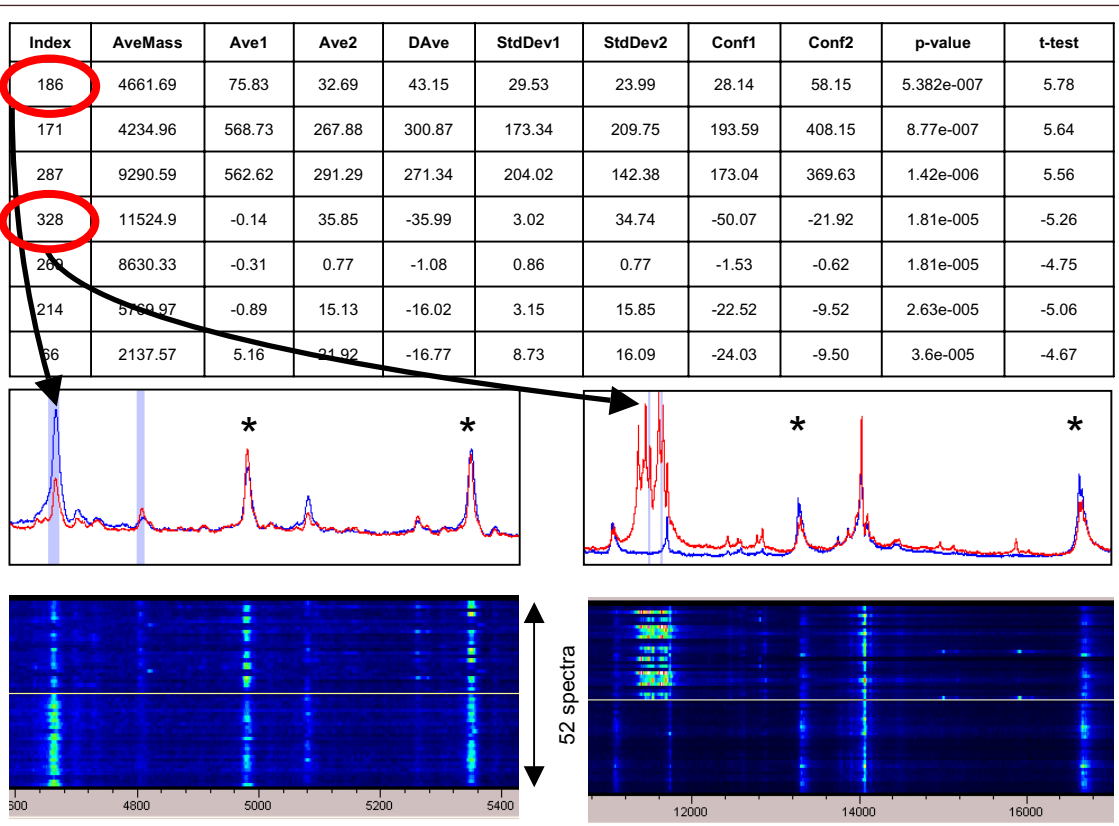
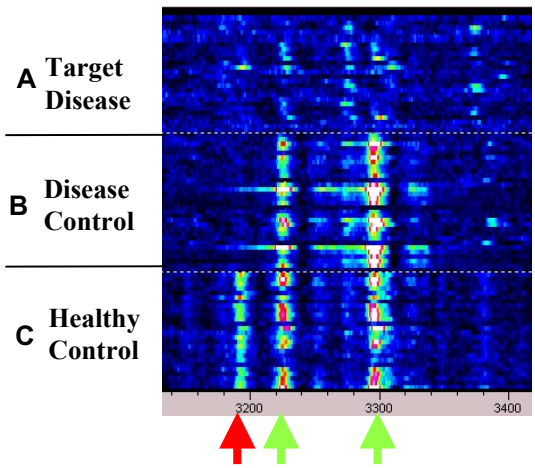


Figure 2. Feedback between visualization and peak statistics. The upper table shows part of the peak statistics results calculated with the ClinProTools software. The peaks are sorted according to decreasing separation power, as indicated by the *P* value. The quality of the proposed peaks can be directly evaluated in the visualization tools. Note the low *P* values showing highly significant differences between the two classes. Signals with nearly identical intensities (marked with asterisks) in the data set can serve as internal controls in comparison to differentially expressed signals. Index, peak number; AveMass, average mass; Ave1+2, average intensity of class 1 and 2; Dave, difference of average intensities; StdDev1+2, standard deviation of class 1 and 2; Conf1+2, confidence interval of class 1 and 2; *P* value, probability that the respective intensity distribution can be observed by chance; *t*-test, result of the *t*-test.

pattern recognition models to analyze independent data to evaluate the quality of these biomarkers to be used in a sample classification for clinical research. Moreover, the statistical approach can be combined with use of the modified genetic algorithm. Here again, the results of both approaches can be directly compared by means of the powerful visualization tools. Peaks used in models generated by the genetic algorithm are highlighted with red or grey bars in the viewer (Figure 2).



C vs. B			B vs. A		
Index	Average Mass	*P* Value[1]	Index	Average Mass	*P* Value[2]
130	4231.77	2.79e-014	164	4232.51	6.47e-011
224	7768.66	2.32e-012	167	4290.00	7.76e-009
126	4114.74	8.57e-012	160	4114.19	4.10e-007
88	2643.59	8.62e-011	245	7769.77	4.60e-007
121	3980.57	8.84e-011	134	3295.21	4.77e-007
125	4077.18	9.41e-011	214	5078.97	7.98e-007
212	6631.86	1.59e-010	130	3225.98	1.59e-006
73	2314.3	1.92e-010	269	9280.27	2.25e-006
102	3192.02	6.01e-010	226	5911.42	3.46e-006
61	2057.36	1.60e-009	255	8133.64	1.27e-005
205	6434.59	2.28e-009	39	1570.54	1.33e-005
230	8136.75	1.57e-008	111	2796.13	1.57e-005
134	4302.54	4.39e-008	155	3978.25	1.71e-005
178	5078.29	5.13e-007	129	3190.07	2.20e-005

Figure 3. Combined analysis of three sample classes (healthy control, disease control, and target disease). The upper panel shows the mass-to-charge ratio (m/z) region of approximately 3100–3400. In this region, biomarker candidates for a general disease condition (approximately 3190 Da, red arrow) and for the target disease (3225 and 3295 Da, green arrows) can be found. The same signals appear also in the statistical peak analysis (see the table in the lower panel). Here, red color indicates peaks with significant *P* values in both disease groups (disease control and target disease) in comparison to healthy controls, while green color indicates peaks with significant *P* values exclusively in target disease in comparison to both control groups.

Currently, the opinion of the scientific community is divided concerning the appropriate approach for biomarker analysis. One major doctrine says that each potential biomarker has to be identified to be of any further clinical use (5). A second doctrine says that a diagnostic pattern, which will give a better classification than existing tests, could be finally applied in the clinic (6). In both cases, an effective visualization of large patient data sets is crucial to get a confident indication of which potential biomarker signals are promising for peptide identification or for future routine application in diagnostic research. Potential biomarker peaks can be selected for in-depth analysis after further enrichment to reduce the sample complexity and analyzed by tandem mass spectrometry (TOF/TOF) analysis. Instruments equipped with a TOF/TOF analyzer, like the UltraFlex used in this study, can be directly used for profiling and for in-depth TOF/TOF analysis. Here, the signal of a potential biomarker is selected in an according mass filter, whereas other signals are discarded. A postacceleration step allows analysis of all metastable fragments of the biomarker candidates (7), which gives a chance for identification in an according database.

Conclusively, ClinProTools is a very powerful software tool for the data interpretation in the field of clinical proteomics. It combines efficient visualization with automated data pretreatment and intuitive statistical analysis. We have used it to determine a number of biomarker candidates in a preliminary data set within a few hours. Last but not least, ClinProTools does not require a deep knowledge of statistics and mathematics to be efficiently used.

ACKNOWLEDGMENTS

The authors would like to thank Mark Flocco for helpful comments.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

1. **Pusch, W., M. Flocco, S.M. Leung, H. Thiele, and M. Kostrzewa.** 2003. Mass spectrometry-based clinical proteomics. Pharmacogenomics *4*:463-476.
2. **Zhang, X., S.M. Leung, C.R. Morris, and M.K. Shigenaga.** 2004. Evaluation of a novel, integrated approach using functionalized magnetic beads, bench-top MALDI-TOF MS with prestructured sample supports, and pattern recognition software for profiling potential biomarkers in human plasma. J. Biomol. Tech. *15*:167-175.
3. **Suckau, D., L. Cornett, and K.O. Kraeuter.** 1998. Automatic acquisition of MALDI-TOF mass spectra. Analysis *26*:M36-M40.
4. **Ransohoff, D.F.** 2004. Rules of evidence for cancer molecular-marker discovery and validation. Nat. Rev. Cancer *4*:309-314.
5. **Diamandis, E.P.** 2004. OvaCheck: doubts voiced soon after publication. Nature *430*:611.
6. **Villanueva, J. and P. Tempst.** 2004. OvaCheck: let's not dismiss the concept. Nature *430*:611.
7. **Suckau, D., A. Resemann, M. Schuerenberg, P. Hufnagel, J. Franzen, and A. Holle.** 2003. A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. Anal. Bioanal. Chem. *376*:952-965.

Address correspondence to:

Wolfgang Pusch
Bruker Daltonik GmbH
Fahrenheitstrasse 4
D-28359 Bremen, Germany
e-mail: WPU@bdal.de