# Effective fishing of characteristic proteome fractions and identification of biomarkers therein:
Application of VisualCockpit to multidimensional chromatogram and MS data

S. Kreusch[1], M. Nagel[3], R. Janetzko[3], G.A. Cumme[1], A. Baum[1], M. Pohl[2], A. Meier-Hellmann[4], H. Rhode[1]

[1]Institute of Biochemistry I, [2]Childrens Hospital, Medical Faculty, Friedrich Schiller University, Jena, [3]Dr. M. Nagel Datenanalyse + Statistik, Oelsnitz, [4]Helios-Clinics, Erfurt, Germany

Dr. Matthias Nagel
Datenanalyse + Statistik
Informationsmanagement
www.dr-mnagel.de · info@dr-mnagel.de
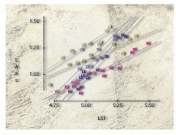
INDRODUCTION

Two different strategies may be used for fishing biomarkers via proteomics based methods :

1) All proteome fractions – eventually some thousands – generated from different samples are analyzed, quantified, and compared. This causes high operating effort and requires high throughput methods for analysis and data handling which are not available with all analytical procedures. This strategy provides the possibility to find all biomarker candidates detectable, however, their number is confined by the analytical methods at hand and the possible operating effort.

2) Otherwise, start with pre-selection of characteristic fractions and focus the operating expense on careful and comprehensive analysis of these fractions using an adequate assortment of different analytical strategies. With this strategy biomarker yield depends on the pre-selection criteria.

Depending on the separation procedure, characteristic fractions may be identified by double staining with 2DE, characteristic tags, other task-specific properties or simply by protein quantification.

Here we introduce an efficient pre-selection based on optical properties of the fractions. Logical selection criteria are applied to sequentially reduce the fractions that have to be analyzed further. Appropriate subsets of fractions are produced and visualized by the software package VisualCockpit. This software can be applied in principle to all data formats and to both fishing strategies, and to evaluation of separation procedures.

Separation Method

A multi-dimensional non-denaturing separation method for intact proteins (Scheme below) according to Horn et al. [Proteomics, 2006, 6, 559] is used. In few words, in the first step (SEC), 0.9 mL serum is separated by a HiLoad Superdex G 200 column (16/60, Pharmacia). The column is equilibrated with 10 mM Tris/HCl, pH 7.4, containing 150 mM NaCl. After the void volume, 1 mL fractions are collected in a 96-deep well micro-plate. Total protein is quantified by UV absorbance measurement at 280 nm [Kreusch et al., Anal. Biochem. 2003, 313, 208]. After parallel dialysis [Horn et al. Patent application: WO 03/049841A1] of aliquots of the 96 fractions obtained, the second separation step (AEC) is performed. Briefly, with ninety-six parallel DEAE-Cellulose columns thirty-five elution steps are performed with collection in 384-well microplates, each micro-plate corresponding to four steps.
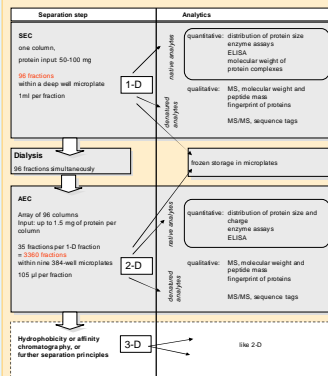
Quantification Method

A method is introduced to evaluate concentration of single proteins within samples using the height sum of all MALDI-MS (matrix-assisted laser desorption-ionization mass spectrometry) peaks that unambiguously match theoretic tryptic peptide masses of the protein sought after. The method uses native chromatographic protein fractionation prior to digestion but does not require any depletion, labeling, derivatization, nor preparation of a compound similar to the analyte. All peak heights of tryptic peptides are normalized with the peak height of a unique standard peptide added to the MALDI-MS samples. The sum of normalized peak heights, Sn, or the normalized mean peak height, Mn, may be used to quantify the respective protein (Bublitz et al. [Proteomics, 2006, 6, 3909]).

VisualCockpit

VisualCockpit provides quick answers to permanently changing questions when making decisions and evaluating multidimensional data regarding large numbers of objects. For routine tasks as, e.g., fitting of model parameters, there are many solutions, only a few, however, for quick answers to ad hoc questions that arise while exploring the data, regarding, e.g., characteristic subsets of objects or interdependence of different data and/or data functions. VisualCockpit serves exactly to this purpose.

The basic principle of VisualCockpit is application of interactive graphics throughout. In contrast to other software as, e.g., Excel, object properties are shown graphically (points, symbols, curves, etc.) instead of indicating numbers. Numerical information can be visualized by hitting selected objects with the cursor. All objects on the VisualCockpit screen are linked together graphically and logically. This makes it possible to visualize high dimensional data sets simultaneously and to analyze them in any depth.

VisualCockpit enables different sights of the data at the same time:
• an aggregate "Management" sight (e.g., which department works with success, which with loss) and
• views of any details (e.g., show me some detail of the costs);
• snapshots - data warehouse - views or
• real time informations.

With VisualCockpit one can extract information from data for deciding on further activity without having to know the data base structure or the data base language.

VisualCockpit is thus used among other things successfully for decision-making on the base of data, e.g., in biomarker search, medicine, and health care.
Moreover, the analysis tools can be used equally well for controlling and reporting tasks.

SUMMARY

Application of VisualCockpit to a multidimensional proteomics approach is shown to facilitate and speed up various biomarker search tasks: Method testing, comparison of thousands of fractions of two and more samples, detection of fractions with total protein level characteristic of clinical state, and fishing of biomarker candidates.

The software package may handle all data types available. It applies logical criteria prescribed by the user to numerical raw data that characterize global protein content of fractions, peptide founds from ESI-MS/MS protein identification, and normalized peak height sums from MALDI analysis that provide concenstration estimates of selected proteins.

RESULTS
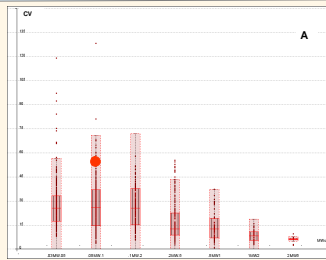
1) Evaluation of the separation method
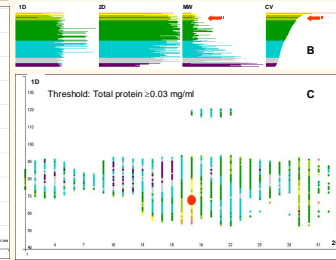
Examples of visualization of linked data in different presentations :

Fig. A: box plot, fig. B: data sort.
Figs. C and D: scatter plots after different filtering

Testing 2-D separation by CV of total protein level:

After six separations of a normal serum sample, mean and CV were determined for each 2-D location. The box plot (Fig. A) shows distributions of CV for selected intervals of mean total protein concentration, the interval x<mean≤y being indicated by xMWy. Points of 2-D matrices (scatter plots, Figs. C, D) are colored according to CV. Combination of box plot with stepwise filtering by selected thresholds of the mean reveal, e.g., that the extremely high CV (marked by •,o, Figs. C, D; marked by ----, Fig B) is correlated with very low protein content and not with methodological error.
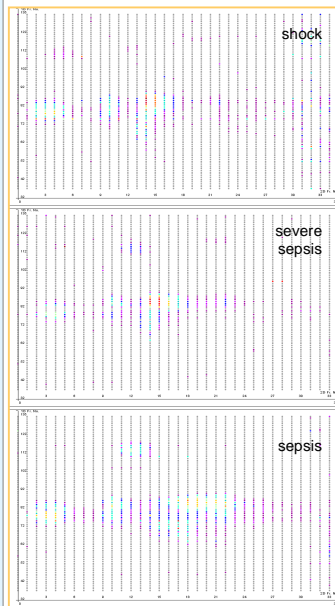
2) Fishing characteristic 2-D fractions

Example 1: One patient with differently severe inflammatory disease: sepsis (seps), severe sepsis (sevr), and septic shock (shoc).
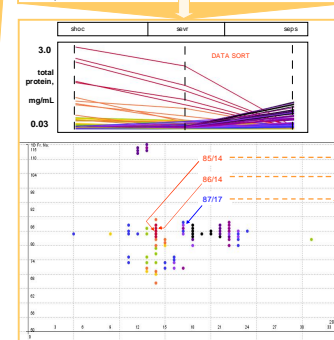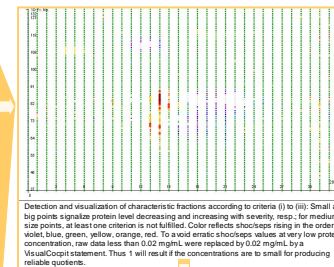
Required: Excel file, columns with 1-D and 2-D fraction numbers and total protein content data from samples.
First line: Variable names, e.g., 1D fract. no., 2D fract. no., seps, sevr, respectively.

Characteristic fractions should fulfill (i) seps ≤ sevr ≤ shoc or vice versa; (ii) seps ≥ 0.03 mg/mL or shoc ≥ 0.03 mg/mL, resp, and (iii) shoc/seps ≥ 2 for reliable measurement.

Detection and visualization of characteristic fractions according to criteria (i) to (iii): Small and big points signalize protein level decreasing and increasing with severity, resp.; for medium size points, at least one criterion is not fulfilled. Color reflects shoc/seps rising in the order violet, blue, green, yellow, orange, red. To avoid erratic shoc/seps values at very low protein concentration, raw data less than 0.02 mg/mL were replaced by 0.02 mg/mL by a VisualCockpit statement. Thus 1 will result if the concentrations are to small for producing reliable quotients.

3) Fishing biomarker candidates

Example 2: Comparison of data of homologous 2-D fractions of serum samples of four probands: three patients with Mb. Alport (full symbols) and one healthy control (empty symbols).

Normalized peak height sum, Sn, and normalized mean height, Mn, correlate with transferrin immunoreactivity better than global protein level.

(Bublitz et al. [Proteomics, 2006, 6, 3909]).

Further evaluation of example 1: Normalized MALDI peak height sum of tryptic peptides of selected biomarker candidates as a function of severity of illness

1D Fr. No	2D Fr. no.	shoc	sevr	seps	shocQseps
85	14	2.33	1.35	0.07	32.01
86	14	2.84	2.20	0.11	26.45
87	17	0.10	0.24	0.41	0.23

Info-table with selected fractions of example 1:

Columns 3 to 5 indicate global protein concentration (mg/mL), column 6 indicates a quotient of protein concentrations

Fraction	Normalized peak height sum (MALDI-MS)			Identified protein entities (LC-ESI-MS/MS)	
	Shock	Severe Sepsis	Sepsis	SwissProtID	Designation
85/14	0.63	0.876	0.01	P02763	ALPHA-1-ACID GLYCOPROTEIN 1
	0.285	0.273	0.043	P19652	ALPHA-1-ACID GLYCOPROTEIN 2
	0.803	1.361	0.262	P01011	ALPHA-1-ANTICHYMOTRYPSIN
	8.798	3.363	0.03	P01009	ALPHA-1-ANTITRYPSIN
	31.373	6.938	0.071	P02768	SERUM ALBUMIN
	3.45	2.713	0.015	P02774	VITAMIN D-BINDING PROTEIN
86/14	0.286	0.018	0.01	P02763	ALPHA-1-ACID GLYCOPROTEIN 1
	1.781	0.39	0.161	P01011	ALPHA-1-ANTICHYMOTRYPSIN
	2.975	0.902	0	P01009	ALPHA-1-ANTITRYPSIN PRECURSOR
	17.284	3.298	0.007	P02768	SERUM ALBUMIN
	0.286	0.423	0.008	P02647	APOLIPOPROTEIN A-I
	3.276	5.344	0.004	P02774	VITAMIN D-BINDING PROTEIN
87/17	0.004	0.046	0.033	P02763	ALPHA-1-ACID GLYCOPROTEIN 1
	0.026	0.116	0.169	P19652	ALPHA-1-ACID GLYCOPROTEIN 2
	0.036	0.147	0.662	P01009	ALPHA-1-ANTITRYPSIN
	1.102	22.001	11.516	P02768	SERUM ALBUMIN

Upon hitting any point within the left diagram with the cursor, VisualCockpit may indicate all parameters related to the respective 2-D fraction, e.g., SEC and AEC (=1-D and 2-D, resp.) fraction numbers and protein concentration ratio, shoc/seps.
Fractions showing strong increase or decrease of concentration with severity of illness according to criteria (i) to (iii) were subjected to stringent ESI-MS/MS protein identification by tryptic peptides. The table presents normalized MALDI peptide peak height sum (Sn) for some proteins found within the three highlighted fractions. Red and blue color indicate increase and decrease with severity, respectively.
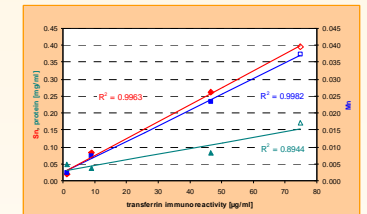
Raw data visualization by VisualCockpit: Protein concentration within 2-D fractions rises in the order gray, violet, blue, green, yellow, orange, red

Extraction of characteristic 2-D fractions according to criteria (I) to (iii): 1 or 0 was assigned to each of the 3360 2-D fractions according to whether or not the above criteria are fulfilled, resp. The subset with 1 (54 and 25 fractions with total protein descending and ascending with severity, resp.) was selected. Upper: Total protein conc. as a function of severity. Lower: Points indicate selected 2-D fractions. Shoc/seps rises in the order black, violet, blue (≤0.5), green (≥2), yellow, orange, red.