

TECHNICAL BRIEF

Multidimensional chromatography: Validation and efficient fishing for biomarkers and fractions containing them using the VisualCockpit software package

Gerhard A. Cumme¹, Stefan Kreuzsch¹, Matthias Nagel² and Heidrun Rhode¹

¹ Institute of Biochemistry I, Medical Faculty, Friedrich Schiller University, Jena, Germany

² Data Analysis & Statistics, Oelsnitz, Germany

2-D native LC yields thousands of fractions especially when applied to sera of different origin. Checking reproducibility of repeated separation of the same serum or searching for biomarker candidates and fractions containing them requires finding, selection, and comparison of interesting data subsets out of huge data volumes. An innovative software package is applied that markedly enhances simplicity, velocity, and reliability of (i) check of reproducibility of the separation method and (ii) analysis of proteomes pertaining to different disease states.

Received: May 9, 2007
Revised: September 25, 2007
Accepted: October 5, 2007

**Keywords:**

Biomarker search / Data mining / Microplate / Multidimensional liquid chromatography / Severe inflammation

Sophisticated bioinformatics tools are available for image analysis with 2-DE [1]; to handle sequence data; to perform protein and peptide alignments; to analyze domain structure, modifications, and interactions [2–5]; to compare mass spectra [6]; and to classify peptide finds according to functional criteria [7]. In addition, open-source systems for controlling workflow and presentation have been described [8]. Nevertheless, the search for biomarkers urgently requires tools [1] to maintain, retrieve, and display large amounts of data so that investigators can keep up with new information, extract the main information, and reliably compare various data sets, especially those of thousands of multidimensional LC fractions [9] originating from different proteomic samples. Although current visualization tools for presenting and

publishing results may seem to be perfect, e.g. [10–12], visualization during data exploration may speed up research and allow timely experimental decisions to be made. Recently we introduced a method that (i) yields native liquid serum fractions and measures the concentration of total protein and single constituents therein and (ii) permits comprehensive protein identification using MS/MS and/or immunoreactivity within the fractions obtained [9, 13]. Bioinformatics tools are needed to evaluate the reproducibility of the separation method and to find marker candidates by quantitatively or qualitatively comparing fractions obtained from samples of different origin. We are testing one such tool, the innovative software package VisualCockpit-Life Science (Data analysis & Statistics, Germany), for its suitability. The package meets important needs: (i) it allows for high-speed data visualization with various types of diagrams according to conditions specified by the user; (ii) subsets of data with features of interest to the user may be highlighted according to criteria specified by the user; (iii) the package accepts Excel and text files and data of different kinds, such as quantifications and annotations; and (iv) subsets of data may be easily selected for further analysis by touching or encircling data

Correspondence: Heidrun Rhode, Institute of Biochemistry, Medical Faculty, Friedrich Schiller University Jena, D-07740 Jena, Germany

E-mail: heidrun.rhode@mti.uni-jena.de

Fax: +49-3641-938612

Abbreviation: CV, coefficients of variation

points with the cursor or by using the calculator and the filter tools found in the package. Optional functions such as measures of reproducibility may be calculated using the calculator tool, displayed, and added as new variables to the data list. All input data such as sample characteristics, fraction numbers, measurement data, and data-processing results are linked, so that touching one data point in one diagram with the cursor highlights that point and its counterparts in all other diagrams present on the screen. The opportunity to visualize data while analyzing them is a large advantage of the software package. A demo version may be visited under www.visualcockpit.de/presentation. To that end version 6 or a later one of the Internet Explorer and the Adobe flash player are required. The flash player can be downloaded from www.adobe.com choosing the tab “solutions and products” and the link “get adobe flash player”. In addition to this technical brief, colored illustrations are available online in the Supporting Information (Figs. A1–A7).

2-D chromatography was validated as follows. A serum sample was separated into 96 1-D SEC fractions in a 96-well microplate [9]. From six repeated separations, fractions from the same well position (homologous fractions) were pooled and mixed, and the mixtures were pipetted into six new microplates, frozen, and stored. All fractions from two such plates were 2-D separated in parallel into 35 AEC fractions each. Two blocks of 96 parallel AEC micro columns (S1 and S2) were used to that end, yielding two series of 35×96 2-D fractions within one working day [14]. This was done three times, and global protein content was determined from UV absorbance in the resulting six series of 2-D fractions [9]. VisualCockpit imports fractional concentration data from an Excel file containing SEC and AEC fraction numbers FrSEC and FrAEC, respectively, and mg/mL protein concentrations S1A1, S2A1 (1st day), S1A2, S2A2 (2nd day), and S1A3, S2A3 (3rd day).

To compare variation between AEC column blocks and between repeated fractionations, intraday and interday coefficients of variation (CV) of total protein concentration are calculated for each 2-D fraction. Formulas used are given in the Supporting Information. Histograms showing the distribution of both these CV values are presented a second after hitting both their names in the list of variables in the VisualCockpit window. After the mean protein concentration for each 2-D fraction is calculated, and colors are assigned to ranges of mean protein concentrations, each bar in the CV histogram is subdivided into colored sections, the lengths of which indicate the number of fractions with the corresponding CV and protein concentration. An example with gray tones is given in Fig. 1; a colored diagram is shown in Fig. A1 (Supporting Information). Analysis is confined to fractions with total protein ≥ 0.02 mg/mL to get meaningful CV values. Figure 1 shows that intraday variation (variation between homologous fractions obtained on the same day by two blocks of 96 parallel AEC microcolumns) is markedly weaker than interday variation (variation between homologous fractions obtained with the same block on different

days). The gray tones indicate that the majority of 2-D fractions contain less than 0.3 mg protein *per* mL.

After CV values for each 2-D fraction are calculated over all six separations, the distribution of these values can be visualized as a function of mean fractional protein concentration by a series of box plots (Fig. A2 in Supporting Information) and simultaneously in Table 1. To this end, a category is assigned to each 2-D fraction corresponding to its mean fractional protein concentration. As expected, high protein concentrations lead to small CV values. The number of 2-D fractions contributing to each category can also be shown in a bar chart or table, and these numbers are included in Table 1. If the SEC fraction number is plotted against the AEC fraction number, each data point in the resulting scatter plot corresponds to one 2-D fraction. Coloring the points according to CV values of protein concentration shows that especially high CV values are found at some AEC fraction numbers. Plotting CV values *versus* mean protein concentration reveals that the high CV values are due not to imperfect performance of these particular AEC elution steps but rather to very low fractional protein concentrations (Fig. A3 in Supporting Information).

To check whether repeated separation can produce different elution profiles, VisualCockpit can simulate shifting among 2-D fractions of different separations by shifting entries within the list of a selected variable, *e.g.*, S2A1, producing a new variable, S2A1s. Relative differences between the entries of shifted and nonshifted variables, $|(S1A1 - S2A1s)/\text{mean}(S1A1, S2A1s)|$, are then calculated as a function of degree of shift. Figure A4 in Supporting Information shows box plots of relative differences obtained after shifting. The smallest differences are obtained by zero shift, indicating that elution profiles are reproducible.

Following this, 2-D fractions containing biomarkers were searched for. The sera of one patient with differently severe states of inflammatory disease, *i.e.*, heavy shock, shock, sepsis, and no criteria indicating sepsis, were 2-D separated. Let C3, C2, C1, and C0 indicate protein content found in homologous fractions belonging to samples with heavy shock, shock, sepsis, and without sepsis criteria. For each quadruple of homologous fractions, the protein concentration quotients Qcrit, *i.e.*, $C3/C2$, $C2/C1$, and $C1/C0$, were calculated together with their minimum and maximum, MinQcrit and MaxQcrit, respectively. A MinQcrit that exceeds a limit above 1 means that total protein concentration is increasing steadily with the severity of disease and *vice versa*. To avoid division by unreliably small concentration values, values below 0.003 mg/mL were replaced by this value (median SD observed near zero concentration). VisualCockpit formed subsets of quadruples with minimum Qcrit ≥ 1.15 and maximum Qcrit ≤ 0.87 to select quadruples with increasing and decreasing protein content, respectively. The selection was confined to quadruples with C1 and C2 ≥ 0.02 mg/mL so that elevated concentrations were in the range of reliable measurement. Selected quadruples can be visualized by (i) plotting maximum Qcrit *versus* minimum Qcrit and

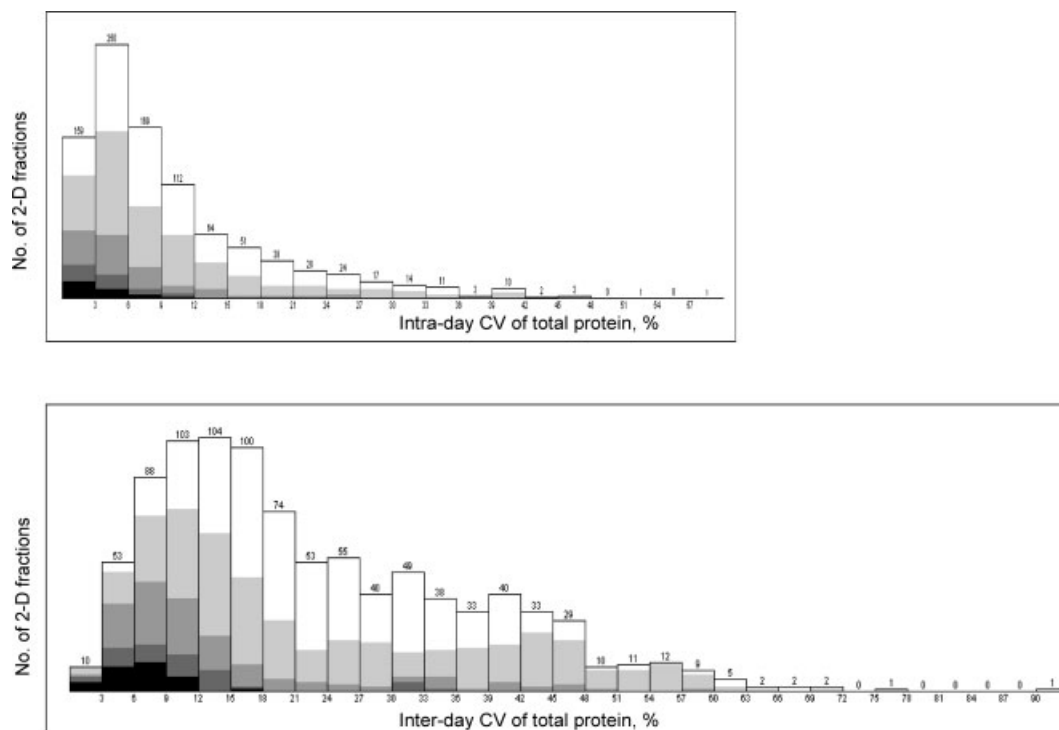


Figure 1. Distribution of intraday and interday CVs in the total protein concentration in chromatographic serum fractions. One sample was separated six-fold by SEC (1-D), and the unique set of pooled homologous fractions was separated six-fold by AEC (2-D). Gray tones indicate ranges of mean total protein concentration. Edges of gray regions correspond to 0.02, 0.07, 0.3, 0.8, 1.4, and 2.7 mg/mL from white to black. Evaluation was confined to 2-D fractions yielding ≥ 0.02 mg protein *per* mL for each of the six separations performed.

Table 1. CV values of fractional protein concentrations: distribution parameters for different ranges of mean protein concentrations

Mean protein range, mg/mL	No. of fracts.	Min. CV	25% pctl.	Median CV	75% pctl.	Max. CV	Mean CV	SD
$0.02 \leq \text{Mean} \leq 0.05$	241	6.31	15.60	21.16	28.59	57.27	22.32	8.87
$0.05 < \text{Mean} \leq 0.1$	288	3.82	14.47	25.30	36.45	80.99	26.52	14.08
$0.1 < \text{Mean} \leq 0.2$	150	5.66	15.08	25.72	37.82	57.89	26.60	12.87
$0.2 < \text{Mean} \leq 0.5$	173	2.71	8.80	12.70	22.66	55.32	17.92	12.81
$0.5 < \text{Mean} \leq 1$	50	0.80	6.94	12.63	19.13	37.24	14.58	9.18
$1 < \text{Mean} \leq 2$	44	2.81	5.43	8.38	10.72	16.01	8.39	3.45
$2 < \text{Mean} \leq 5$	11	2.52	6.08	6.25	6.77	9.55	6.12	1.96

CV, %; pctl., percentile; *cf.* also Supporting Information, Fig. A2.

(ii) plotting the SEC fraction number *versus* the AEC fraction number. An example is presented in Fig. 2; a colored diagram is shown in Fig. A5 (Supporting Information). The limits chosen for Q_{crit} correspond to at least a 15% increase or decrease, although Table 1 shows some higher CV values. Such limits permit detection of quadruples exhibiting strong variation between neighboring disease states, see, *e.g.*, the fat points in Fig. 2. Their ordinates indicate at least one concentration pair with a ratio ≥ 2 , which is well out of the CV ranges shown in Table 1.

Candidate biomarkers were then identified as follows. Three sera obtained from another patient during decreasing

severity of sepsis, *i.e.*, septic shock, severe sepsis, and sepsis, were 2-D separated and the total fractional protein concentrations were determined. Homologous fractions whose protein levels varied in a clear-cut manner according to severity of disease were selected using VisualCockpit. To demonstrate the search strategy, selected fractions were digested by trypsin and thereafter analyzed by LC-ESI-MS/MS [9] 2 ± 0.7 times (mean \pm SD). For each protein identified, the mean number of its tryptic peptide founds was added up overall selected 2-D fractions for each disease state. As former experiments showed that the number of peptide

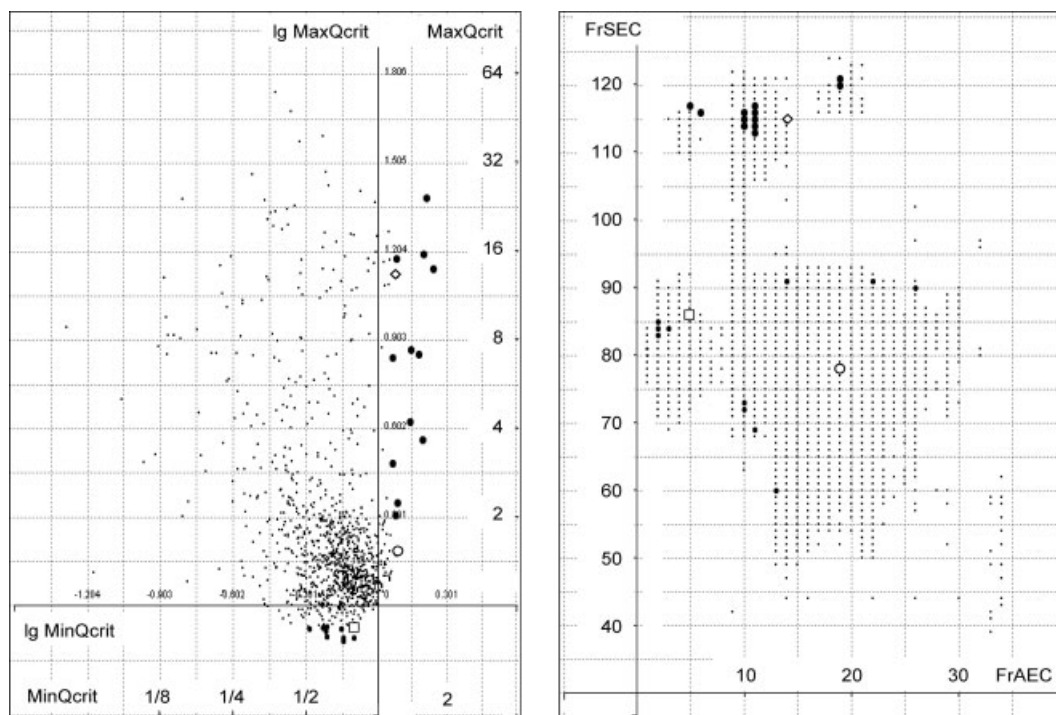


Figure 2. Search for 2-D fractions that may contain sepsis markers. Sera of one patient with differently severe states of inflammatory disease, *i.e.*, heavy shock, shock, sepsis, and no criteria indicating sepsis, were separated by SEC (1-D) and AEC (2-D). For each quadruple of homologous fractions, the protein concentration quotients Qcrit, *i.e.*, C3/C2, C2/C1, and C1/C0, were calculated together with both their maximum and minimum values, MaxQcrit and MinQcrit, respectively. Left diagram: outer ordinate and abscissa, MaxQcrit and MinQcrit; inner ordinate and abscissa, log(MaxQcrit) and log(MinQcrit), respectively. Fat points indicate the concentration increases with the severity of disease (MinQcrit ≥ 1.15), medium points indicate the opposite (MaxQcrit ≤ 0.87). Size of points was adjusted by VisualCockpit. Points within the second quadrant indicate fractions with nonunidirectional variation of concentration according to the severity of disease. Right diagram: scatter plot FrSEC *versus* FrAEC. Ordinate, eluted SEC volume, mL; abscissa, and AEC fraction number. Open circles, squares, and rhombuses illustrate how data points are linked within both VisualCockpit diagrams.

found increased with the concentration of the respective protein, founds were assumed to be an estimate of concentration. Figure A6 in the Supporting Information shows a bar chart in which each bar belongs to a protein and its subdivision into red, yellow, and green parts indicates the peptide founds pertaining to septic shock, severe sepsis, and sepsis. Proteins with peptide founds that clearly depend on severity of illness may be assumed to be marker candidates. After a candidate is selected by hitting its bar with the cursor, its distribution over the selected 2-D fractions is illustrated with a scatter plot of SEC *versus* AEC fraction numbers. Figure A7 in the Supporting Information shows an example.

Combining multidimensional separation with MS/MS protein identification and MS based measures of protein concentration may thus indicate protein modifications and/or protein isoforms that cause a modified protein or its isoform to elute at an altered position in the 2-D chromatogram. Besides concentration measures, various other input data may be obtained depending on analytical task, sample pretreatment, and analytical tools, *e.g.*, data concerning spectral characteristics, enzyme activities, immunoreactivities, peptide composition, alteration after treat-

ment with enzymes that undo PTMs, *etc.* All these data may be analyzed as a whole to compare samples of different origin.

We thank Mrs. Helga Endmann and Mrs. Bärbel Tautkus for excellent technical assistance. The financial support of BMBF (PTJ-Bio/0312699) and DFG (Ho 1311/3-1) is gratefully acknowledged.

The authors have declared no conflict of interest.

References

- [1] Palagi, P. M., Hernandez, P., Walther, D., Appel, R. D., Proteome informatics: Bioinformatics tools for processing experimental data. *Proteomics* 2006, 6, 5435–5444.
- [2] Suresh, S., Mohan, S. S., Mishra, G., Hanumanthu, G. R. *et al.*, Proteomic resources: Integrating biomedical information in humans. *Gene* 2005, 364, 13–18.

- [3] Lu, H., Shi, B., Wu, G., Zhang, Y. *et al.*, Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochem. Biophys. Res. Commun.* 2006, **345**, 302–309.
- [4] Li, D., Gao, W., Ling, C. X., Wang, X. *et al.*, IndexToolkit: An open source toolbox to index protein databases for high-throughput proteomics. *Bioinformatics* 2006, **22**, 2572–2573.
- [5] America, A. H. P., Cordewener, J. H. G., van Geffen, M. H. A., Lommen, A. *et al.*, Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. *Proteomics* 2006, **6**, 641–653.
- [6] Ketterlinus, R., Hsieh, S., Teng, S., Lee, H., Pusch, W., Fishing for biomarkers: Analyzing mass spectrometry data with the new ClinProTools™ software. *BioTechniques* 2005, **38**, S37–S40.
- [7] Graham, R. L. J., O'Loughlin, S. N., Pollock, C. E., Ternan, N. G. *et al.*, A combined shotgun and multidimensional proteomic analysis of the insoluble subproteome of the obligate thermophile, *Geobacillus thermoleovorans* T80. *J. Proteome Res.* 2006, **5**, 2465–2473.
- [8] Rauch, A., Bellew, M., Eng, J., Fitzgibbon, M. *et al.*, Computational proteomics analysis system (CPAS): An extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* 2006, **5**, 112–121.
- [9] Horn, A., Kreuzsch, S., Bublitz, R., Hoppe, H. *et al.*, Multidimensional proteomics of human serum using parallel chromatography of native constituents and microplate technology. *Proteomics* 2006, **6**, 559–570.
- [10] Jones, A., Faldas, A., Foucher, A., Hunt, E. *et al.*, Visualisation and analysis of proteomic data from the procyclic form of *Trypanosoma brucei*. *Proteomics* 2006, **6**, 259–267.
- [11] Morisawa, H., Hirota, M., Toda, T., Development of an open source laboratory information management system for 2-D gel electrophoresis-based proteomics workflow. *BMC Bioinformatics* 2006, **7**, 430.
- [12] Linsen, L., Löcherbach, J., Berth, M., Becher, D., Bernhardt, J., Visual analysis of gel-free proteome data. *IEEE Trans. Vis. Comput. Graph.* 2006, **12**, 497–508.
- [13] Bublitz, R., Kreuzsch, S., Ditzel, G., Schulze, M. *et al.*, Robust protein quantitation in chromatographic fractions using MALDI-MS of tryptic peptides. *Proteomics* 2006, **6**, 3909–3917.
- [14] Rhode, H., Kreuzsch, S., Cumme, G. A., Baum, A. *et al.*, Search for serum biomarkers using an approach based on native multidimensional fractionation. *Shock* 2006, **26**, (Suppl. 1), 1.