EDITORS' CHOICE

BIOTECHNOLOGY and BIOENGINEERING

Accurate Retention Time Determination of Co-Eluting Proteins in Analytical Chromatography by Means of Spectral Data

Florian Dismer, Sigrid Hansen, Stefan Alexander Oelmeier, Jürgen Hubbuch

Chair for Biomolecular Separation Engineering, Institute of Engineering in Life Sciences,
Karlsruhe Institute of Technology, Engler-Bunte Ring 1, Karlsruhe 76131, Germany;
telephone: 49-721-60846236; fax: 49-721-60846240; e-mail: florian.dismer@kit.edu

ABSTRACT: Chromatography is the method of choice for the separation of proteins, at both analytical and preparative scale. Orthogonal purification strategies for industrial use can easily be implemented by combining different modes of adsorption. Nevertheless, with flexibility comes the freedom of choice and optimal conditions for consecutive steps need to be identified in a robust and reproducible fashion. One way to address this issue is the use of mathematical models that allow for an in silico process optimization. Although this has been shown to work, model parameter estimation for complex feedstocks becomes the bottleneck in process development. An integral part of parameter assessment is the accurate measurement of retention times in a series of isocratic or gradient elution experiments. As high-resolution analytics that can differentiate between proteins are often not readily available, pure protein is mandatory for parameter determination. In this work, we present an approach that has the potential to solve this problem. Based on the uniqueness of UV absorption spectra of proteins, we were able to accurately measure retention times in systems of up to four co-eluting compounds. The presented approach is calibration-free, meaning that prior knowledge of pure component absorption spectra is not required. Actually, pure protein spectra can be determined from co-eluting proteins as part of the methodology. The approach was tested for size-exclusion chromatograms of 38 mixtures of co-eluting proteins. Retention times were determined with an average error of 0.6 s (1.6% of average peak width), approximated and measured pure component spectra showed an average coefficient of correlation of 0.992.

Biotechnol. Bioeng. 2013;110: 683–693.

KEYWORDS: chromatography; retention time determination; overlapping peaks; UV absorption spectra; principal component analysis; co-elution

Correspondence to: F. Dismer
Additional supporting information may be found in the online version of this article.
Received 19 July 2012; Revision received 14 September 2012;
Accepted 26 September 2012
Accepted manuscript online 5 October 2012;
Article first published online 23 October 2012 in Wiley Online Library
(http://onlinelibrary.wiley.com/doi/10.1002/bit.24738/abstract)
DOI 10.1002/bit.24738

Introduction

Chromatographic separation of proteins has long been one of most often applied techniques in the field of biotechnology when it comes to protein analytics and preparative separation. Especially in the field of analytical chromatography for small molecules a lot of effort has been put into developing new chromatographic materials and devices (monolithic columns, fused silica columns, HT-UHPLC, etc.) to drive peak resolution, peak capacity, and throughput to the max (Guillarme et al., 2010).

In most analytical cases, baseline separation of all analytes is desirable in order to allow for an accurate and robust determination of retentions times and peak areas (Piaggio et al., 2005). According to Giddings (Giddings, 1967) the peak capacity of a column can be calculated by to the following equation:

$$n_c = 1 + \frac{\sqrt{N}}{4R_S}\ln\left(\frac{k_l + 1}{k_f + 1}\right) \tag{1}$$

where N is the number of theoretical plates, R_S is the resolution and k_l and k_f are the retention factors for the first and last eluting component. According to this equation, a typical analytical gel filtration column suitable for protein analytics (e.g., a Superdex column by GE Healthcare, 30 cm, 24 mL bed volume, 30,000 N/m) has a peak capacity of approximately eight for a target resolution of 1.5, meaning that it can resolve a maximum of eight baseline-separated peaks. As soon as peaks start to overlap, peak maxima can be shifted and peak de-convolution becomes mandatory, even if distinct maxima can still be observed (Berthod, 1991; Foley and Dorsey, 1983; Hanggi and Carr, 1985; Lan and Jorgenson, 2001; Naish and Hartwell, 1988). However, it is an inherent feature of all de-convolution approaches that they cannot necessarily determine the accurate number of components in a peak when resolution is very poor. If one could obtain this number and additionally get reasonable starting parameters for peak fitting of all components, peak capacities could be increased tremendously. Reducing the resolution R_S in Equation (1) from 1.5 to 0.1 would increase the maximum number of peaks that could be resolved from 8 to 100.

Other than in analytical chromatography, a high level of resolution is often not necessary in preparative chromatography as the goal is to reduce feed complexity rather than identifying and quantifying all individual impurities. Nevertheless, as predictive modeling of preparative chromatography as a tool for process development became increasingly popular over the past decade (Coffman et al., 2008; Liapis, 1990; Osberghaus et al., 2012a,b,c; Tejeda-Mansir et al., 2001; Wright et al., 1998), chromatographic determination of model parameters is required to have a high level of accuracy. Additionally, parameters are needed for many components resulting in retention time measurements at different salt concentrations under isocratic conditions or with different salt gradient slopes for all components of interest. This is usually done in single-component injections, which can result in tedious work as pure components are often not readily available.

One way to address this issue is to apply multidimensional chromatography to reduce system complexity as nicely shown by Ahamed et al. (2009). They applied a pH gradient on an ion-exchange column in the first dimension followed by linear salt gradients with varying slopes for SMA parameter determination (steric mass action model according to Brooks and Cramer (1992)) in the second dimension. As proteins are usually not baseline separated in all gradient elution runs needed for SMA parameter determination, additional analytics are necessary. For this purpose, SDS page gel electrophoresis was used for fractions collected in the second dimension to determine retention times for single components that co-elute. Although this allows for parameter determination of rather complex mixtures, standard gel electrophoresis is time consuming and thus limits the number of fractions which can be analyzed within a reasonable amount of time. With that the accuracy of retention time determination is limited. Assuming a number of 20 fractions in the first and in the second dimension and a total of three different gradient slopes in the second dimension (which is the minimum for SMA parameter determination) one would end up analyzing $20 \times 20 \times 3 = 1{,}200$ fractions. The accuracy would be about 1/20 of the gradient volume. For a gradient length of 10 column volumes (CV) this would translate into an accuracy of 0.5 CV.

Parallel to developing new hardware, ‘‘self-modeling curve resolution’’ was introduced in the 1970s (Bu and Brown, 2000; Lawton and Sylvestre, 1971; Osten and Kowalski, 1984) that allows for the determination of retention times and peak areas of co-eluting small molecules that greatly differ in their UV absorption spectra. A review by Guillarme et al. (2010) gives a comprehensive overview of current progress in the field. This approach has been pushed forward and was recently used to identify compounds in a 2D liquid chromatographic analysis of small molecules in complex samples (Bailey and Rutan, 2011).

Recent advances in protein analytics by Hansen et al. (2011) have shown that selective protein quantification in samples consisting of up to three components can accurately be done by means of UV absorption sum spectrum of the mixture. This nicely shows that UV absorption spectra of proteins can be used as a unique ‘‘fingerprint’’ highlighting the potential of fast, non-invasive photometric assays. Nevertheless, the major limitation of this approach is that it needs to be calibrated with pure component spectra or at least with samples of known composition; it is thus not applicable to samples of unknown composition. And despite the fact that multivariate curve resolution techniques have widely been applied to small molecules, the awareness of its potential in biotechnological applications is limited. To our knowledge the only publication including an example of co-eluting proteins, was published in 1985 by Vandeginste et al. (1985) and co-elution in general has only been investigated for up to three small molecules. Compared to small molecules, absorption spectra of proteins often share a high degree of similarity as only three different amino acids are essentially responsible for spectra differences in the range of 260–300 nm. Additionally all three amino acids are usually simultaneously present in a protein and their spectra show a significant overlap.

In the following, we exploit differences in UV absorption spectra between proteins to accurately determine retention times of proteins and peak areas of co-eluting proteins. The methodology used does not require calibration, thus samples of unknown composition can be analyzed. We chose a series of 2-, 3-, and 4-component injections onto a size-exclusion column, with resulting chromatograms all showing only one distinct maximum to systematically study the possibilities and limitations of this approach. For all investigated systems, we could resolve retention times with an average error of 0.6 s which was about 1.6% of the average peak width of all components. The average error was independent of the number of components injected; it is thus to be expected that the presented approach is also applicable to more complex samples. Additionally we were able to approximate pure-component spectra with relatively high accuracy in most cases, allowing for an identification of co-eluting proteins when a database with pure-component spectra is available.

Materials and Methods

Proteins

The proteins used are summarized in Table I. Stock solutions were prepared at a concentration between 0.2 and 1.0 g/L. For all proteins, the concentration was adjusted to give absorption spectra of similar intensity in the range from 240 to 300 nm.

Table I. List of proteins used, including data from single component injections.

Protein	Retention time (min)	Peak area (mAU_{280} mL)
mAb 4	6.672	2.82
Glucose oxidase	6.770	12.05
mAb 3	6.841	5.17
Catalase	6.883	5.21
mAb 1	6.970	4.54
mAb 2	7.050	6.79
HSA	7.401	8.18
Avidin	7.451	11.05
Ovomucoid	7.758	6.67
Ovalbumin	7.833	7.53
β-Lactoglobulin	7.974	10.83
Hemoglobin, human	8.046	8.13
Carbonic anhydrase	8.269	11.74
Myoglobin	8.475	6.52
α-Chymtrypsinogen	8.529	8.95
α-Lactalbumin	8.540	11.91
Cytochrome *c*	8.615	10.11
Thaumatin	8.745	9.12
Ribonuclease A	8.747	11.02
Lysozyme, human	9.090	8.82
Lysozyme, chicken	9.197	10.18
Subtilisin	9.202	7.74

SEC Runs

All SEC runs were performed on a UltiMate3000 RSLC 2× Dual System from Dionex (Sunnyvale, CA) together with a Zenix SEC-300 (4.6 mm × 300 mm) column at a flow rate of 0.4 mL/min. Protein stock solutions were injected with a volume of 5 μL. UV absorption spectra were measured in the range of 240–300 nm with 1 nm spacing every 40 ms. Data were exported with the software Chromeleon® (6.80 SR10) in text file format.

Data Handling and Conditioning

All data handling was done with Matlab2011a (The Mathworks Natick, ME). All 2D-chromatograms (UV signal over time and wavelength) used for validation of retention time calculations are "virtual" chromatograms that were generated from single component runs by adding up the UV absorption signals. It should be noted here that this potentially increases the level of noise present in the virtual chromatograms. For that reason final data was mildly smoothed to reduce noise over time using the *csaps* function of Matlab (cubic smoothing spline function with $P = 0.99995$, for detail please refer to the Matlab manual). By adjusting the smoothing strength P, the sensitivity of the following spectral analysis can be fine-tuned. The sensitivity of peak recognition is maximal for $P = 1$ (no smoothing) but also more prone to noise.

Principal Component Analysis

A principal component analysis can generally be used to remove redundancy and reduce complexity in a dataset by projecting the data onto a number of principal components. If for example a dataset includes 10 absorption spectra measured between 260–300 nm (in 1 nm steps) of the same protein measured at different concentrations, the complete dataset would consist of $41 \times 10 = 410$ data points. A PCA would give a vector of loadings for the first PC that has the same shape as the absorption spectrum. All 10 spectra are then linear combinations of the loadings vector and a factor for each concentration. In this way the size of the dataset is reduced to $41 + 10 = 51$ data points without losing information. For this example one PC is sufficient to capture all information.

If the 10 spectra were measured for different mixing ratios of two proteins, the PCA would need two PCs to describe the information inherent in the data. The first PC always captures the highest degree of variation. In this example the loadings vector would be similar to the average of both pure component spectra and would capture maybe 90% of the variation in the dataset, depending on the mixing ratios and the spectral difference of the two proteins used. The second PC would be similar to the difference between both pure component spectra and would ideally capture the remaining 10% of the variation. Each of the 10 spectra would be a linear combination of a factor A times the loadings vector for the first PC and a factor B times the loadings vector for the second PC. In this case the dataset would be reduced to $2 \times 41 + 2 \times 10 = 102$ data points.

In the approach presented here, we use the PCA only to determine the amount of variation in a set of spectra. The only information which is used form the PCA is the variation captured by the first PC. Our approach is similar to a "Fixed Size Moving Window-Evolving Factor Analysis" which is described in detail for example in (Lawson and Hanson, 1974). We used a moving window of fixed size of 10 consecutive normalized spectra, resulting in a time frame of $10\,\text{ms} \times 40\,\text{ms} = 400\,\text{ms}$. Principal component analysis was performed using the *princomp* function of Matlab. The variance σ captured by the first principal component was calculated and used to determine the number of components present in an elution peak as described in the following. In the plots presented later, $1 - \sigma$ is plotted for reasons of clarity.

Pure Component Spectra: Initial Guess

Initial guesses of pure component spectra were extracted whenever $1 - \sigma$ plotted over time reached a minimum. Figure 1 shows exemplary data. With only one component (Fig. 1A: ribonuclease A) the minimum always coincides with the peak maximum as noise decreases with increasing signal intensity. At this point basically all 10 normalized spectra are identical. Figure 2 shows a co-elution of two proteins: glucose oxidase and catalase. Both proteins elute with a difference of 6.8 s. The initial guesses of the pure component spectra are the spectra measured at the minima of $1 - \sigma$.

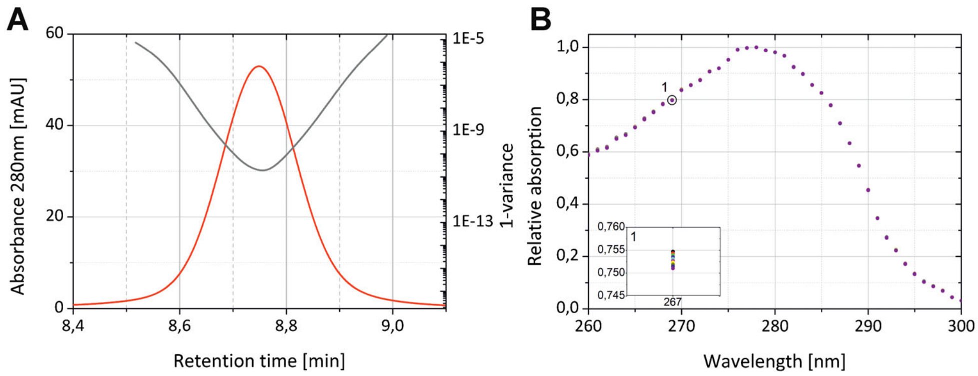

Figure 1. UV trace at 280 nm of a single injection of pure ribonuclease A (**A**) The value plotted on the y-axis is calculated as $1 - \sigma$ with σ being the variance captured in the first principal component of the spectra analyzed. **B**: Ten normalized spectra taken from the early peak as an exemplary dataset for PCA. The PCA is used to calculate the variance this dataset. Subfigure 1 shows a zoom on data measured at 267 nm to show the degree of variance.

Retention Times: Initial Guess

Each measured spectrum throughout the elution peak is a linear combination of the pure component spectra. Using the initial guesses for the pure spectra (see ‘Pure Component Spectra: Initial Guess’ section) one can solve the following quation to get initial elution profiles of the components. The equation is solved for each spectrum through the entire elution peak:

$$\begin{pmatrix} A_{\lambda 1} \\ A_{\lambda 2} \\ \vdots \\ A_{\lambda n} \end{pmatrix} = \sum c_i^t \begin{pmatrix} \varepsilon_{\lambda 1} \\ \varepsilon_{\lambda 2} \\ \vdots \\ \varepsilon_{\lambda n} \end{pmatrix}_i \tag{2}$$

where A_λ is the absorption at a wavelength λ, ε_λ the relative extinction coefficient at a wavelength λ and c_i^t the concentration of the component i at time t. This set of linear equations is solved using the *lsqnonneg* command of Matlab that is based on a publication by Lawson et al. (Lawson and Hanson, 1974). The resulting component specific data can then be fitted with a Gaussian function. More complex functions (e.g., an exponentially modified Gaussian function) can also be used as discussed elsewhere (Caballero et al., 2002). Since size exclusion chromatography was used here, elution peaks were symmetrical and a Gaussian function was sufficient.

Spectra and Retention Time Refinement

Up to this point we have an initial guess for pure-component spectra and an initial guess for the elution profiles. A Gaussian function can then be used to calculate more accurate concentration profiles and with this information, Equation (2) can be solved in the same way as described in the previous section to determine more accurate pure component spectra (ε_λ) for all components. In this case there is one unknown parameter (wavelength-specific extinction coefficient) for each component and wavelength. For each unknown parameter there are several hundred measured absorption values, depending on the width of the elution peak. Equation (2) with either c_i^t or ε_λ as unknowns can then be solved repeatedly until concentration profiles and pure component spectra do not change anymore. In most case, about 15–25 iterations were sufficient. Finally, the concentration profiles can be integrated to give peak areas.

Quality of the Results

To evaluate the quality of the procedure, the determined pure component spectra were compared to spectra from single component injections. Both spectra were fitted linearly and the resulting root-mean-square errors were calculated. Further, retention times and peak areas were directly compared to results obtained from single injections.

Results and Discussion

A table with a detailed overview of all results (including system composition, chromatographic resolution in these systems, accuracy of retention time, and peak area determination and quality of approximated pure component spectra) is part of the Supplementary Material available online.

1-Component Systems

The presented approach is based on the detection of changes in UV absorption spectra measured at the column outlet over time. Ideally, if a single pure protein is injected, the normalized UV absorption spectra of the eluting protein measured over time are essentially identical. In this context

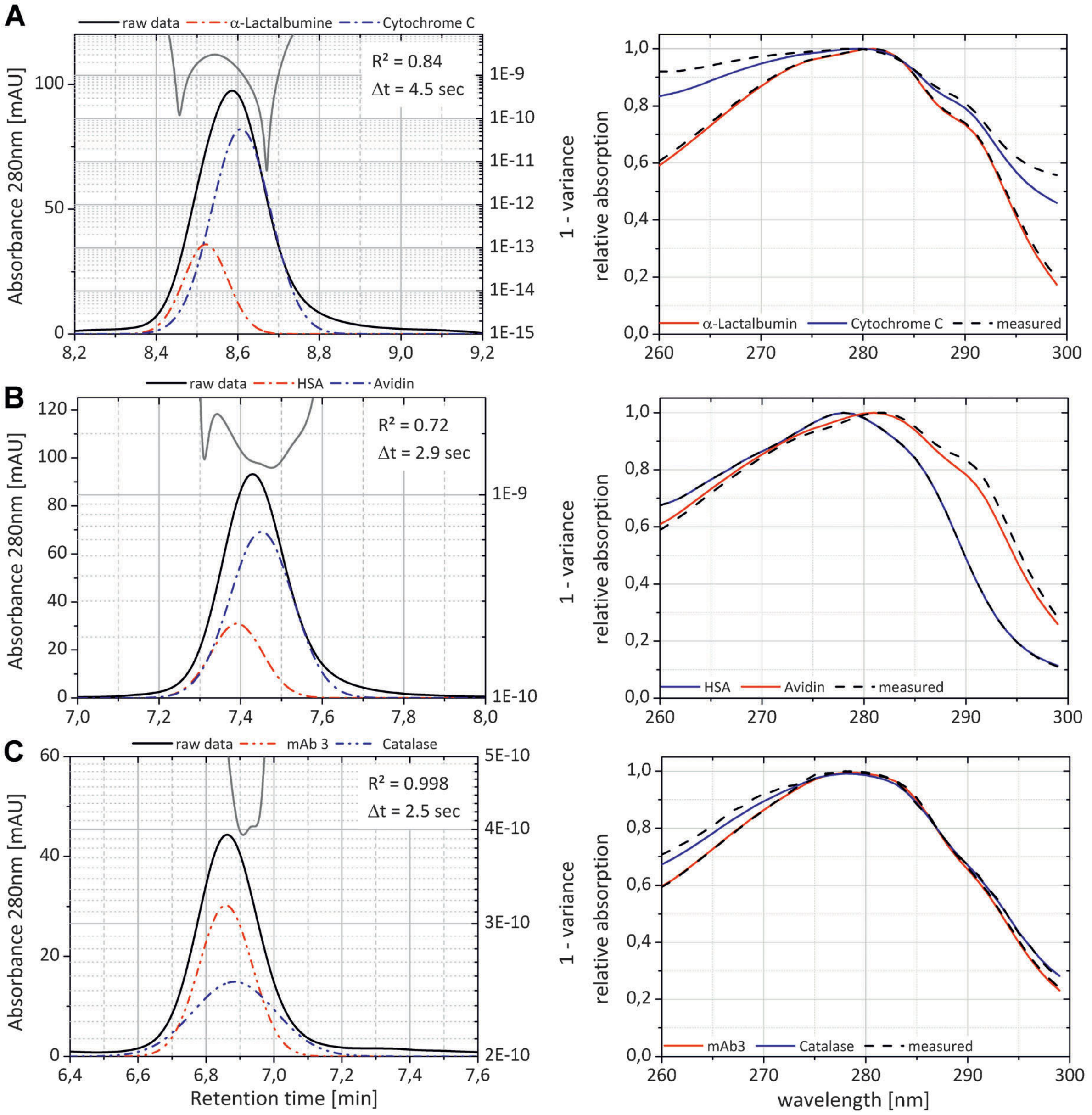

Figure 2. **Left**: Elution peaks of three different two component systems (solid lines). Dashed lines show single component peaks after fitting and optimization. Gray lines show 1-variance of the first principal component analysis. **Right**: Extracted pure component spectra of both components (dashed lines). Solid lines show the reference spectra. **A**: α-Lactalbumine & cytochrome *c*, (**B**) Human serum albumin & avidin, (**C**) mAb3 & catalase. R^2 values indicate similarity of pure component spectra, Δt values the elution time differences from single component runs.

"ideal" means that (1) the measurement is free of noise and (2) measurement of absorption spectra is in the linear range of the detector.

If more than one component elutes, spectra measured over time are no longer identical and spectral variation can be detected in multiple ways. The simplest approach would be to fit two consecutive spectra with a linear equation. In this case, R^2 would be an indicator for the similarity of the two spectra. If one wants to compare more than two spectra, a principal component analysis (PCA) would be a suitable tool.

PCA maps the information inherent in the spectra onto principal components in a way that maximizes the variance captured by a minimal number of principal components. For a number of identical spectra, all information can be mapped onto the first principal component, it thus accounts for all variation present in the data. If the spectra are not identical, more than one component is needed. The variance captured by the first component can thus act as an indicator for the similarity of spectra.

Figure 1 shows the variance (σ) captured by the first PC for the injection of pure ribonuclease A. For reasons of clarity $1-\sigma$ is plotted. As described above, for the ideal case one would expect a straight line as all normalized spectra are identical. For real data, $1-\sigma$ reveals a minimum that coincides with the elution peak maximum, as detector noise decreases with increasing signal intensity. Due to a normalization of the spectra, the absolute noise significantly increases with decreasing signal intensity. This noise adds variation to the data that cannot be captured by the first principal component as it varies from spectrum to spectrum (random noise). As a result, σ reaches a maximum for high signal intensities, and $1-\sigma$ shows a minimum. It should be noted that the variance not captured by the first PC is only about 10^{-9}% in this example.

2-Component Systems

If two components co-elute but to some extent differ in their retention time, the minima observed for $1-\sigma$ over time will no longer coincide with the peak maximum. One of two scenarios can then occur: (1) Minima can be found where only one of the two components elutes, that is, at the very beginning or the very end of the peak. In this case the minimum can be found where absorption of the pure component is highest and thus noise is lowest. This can be seen in Figure 2A and B. If both peaks completely overlap, none of them elutes as a pure component. In this case minima coincide with the maxima of each component (see Fig. 2C). At the peak maximum the concentration of one component is more or less constant over time, while the concentration of the other component is not, thus spectra changes are minimal. In both cases, the number of minima correlated with the number of components. Therefore, initial guesses for pure component spectra can be extracted at the position of the minima. Those can be used to iteratively determine pure component spectra and retention times of the individual components as described in Materials and Methods Section.

To test the robustness of this approach, a total number of 18 2-component systems was investigated by comparing retention times and pure component spectra from single injections with the results generated with our approach. Peak resolution in these systems was in the range of 0.008 (human lysozyme & subtilisin, $\Delta t_r = 0.3$ s) to 0.268 (cytochrome *c* & ribonuclease A, $\Delta t_r = 7.9$ s). Components in these systems co-eluted in one peak showing only one distinct maximum. It should be mentioned again that multi-component chromatograms discussed in the following, are virtual chromatograms generated from single injection runs.

Three representative examples are shown in Figure 2. The first system shown (Fig. 2A) is the co-elution of α-lactalbumin and cytochrome *c*, both eluting with a difference in retention time of 4.5 s according to single-component injections. Injected as a 2-component system, the peak resolution would be 0.15. When plotted over time, $1-\sigma$ of the first principal component showed two distinct minima indicating two species. After refinement of pure component spectra, concentration traces of both components could be calculated and elution peaks were fitted using a Gaussian function (for details refer to Materials and Methods Section). The determined retention times had an accuracy of 1.2 s for α-lactalbumin and 0.2 s for cytochrome *c* compared to retention times from single injection runs. The pure component spectra had a root-mean-square error ($\text{RMSE} = 1 - R^2$) of 0.002 and 0.034 respectively when correlated with the true spectra measured in single-component runs.

The second example is the co-elution of human serum albumin and avidin (Fig. 2B). According to single-component injections they elute with a time difference of 2.9 s, translating into a peak resolution of 0.09 for a two component run. The difference in their absorption spectra is more pronounced (indicated by the lower R^2 value on the top right corner of the plot). Again, plotting the variance from the principal component analysis revealed two minima indicating two components in the elution peak. Their retention time could be determined with an accuracy of 0.5 s for HSA and 0.8 s for avidin. The RMSE values of the extracted pure component spectra were 0.003 and 0.025, respectively.

The third example shows the co-elution of mAb3 and catalase with an elution time difference of 2.5 s and a peak resolution of 0.07. This example was challenging as the spectra of both components are very similar (indicated by an R^2 of 0.998 for their correlation). Nevertheless, elution times could be determined with an accuracy of 0.5 and 0.1 s. The extracted spectra had RMSE values of 0.002 and 0.009 compared to the real pure component spectra.

For all 18 systems investigated, an average accuracy of the determined retention times of 0.60 s was achieved. The average RMSE value for the determined pure component spectra was 0.010.

3-Component Systems

To show that the presented approach also works for more complex systems, we investigated a total number of thirteen 3-component systems. The difference in retention time between first and last eluting component was in the range of 8.4–16.8 s. Three representative examples are shown in Figure 4. Again, the number of components correlates with by the number of minima when $1-\sigma$ is plotted over time. For the first and last eluting component, the minima coincided with pure component elution at the highest absorption signal possible (and thus lowest noise). For the component eluting in the middle, minima roughly coincided with the concentration maxima of this component.

For the first example (Fig. 4A: glucose oxidase, catalase, & mAb2), retention times had an accuracy of 0.4, 0.2, and 0.4 s, extracted pure component spectra had an RMSE of 0.025, 0.002, and 0.005. Of all 13 systems investigated, the retention time differences were highest in this example (16.8 s from the first to the third component).

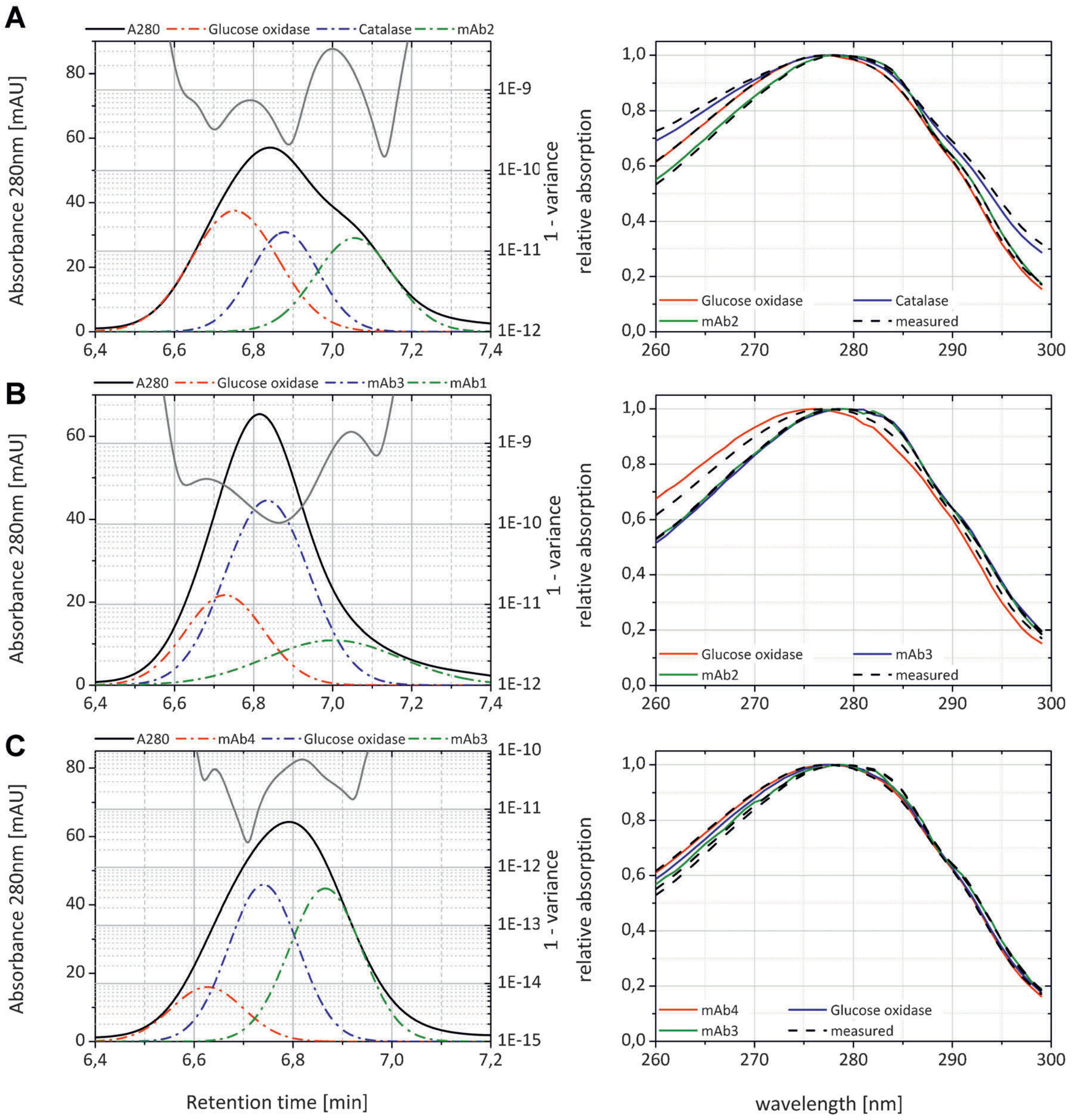

Figure 3. **Left**: Elution peaks of three different three component systems (solid lines). Dashed lines show single component peaks after fitting and optimization. Gray lines show 1-variance of the first principal component analysis. **Right**: Extracted pure component spectra of all components (dashed lines). Solid lines show the reference spectra. **A**: Glucose oxidase, catalase & mAb2; (**B**) Glucose oxidase, mAb3, & mAb1 (**C**) mAb4 Glucose oxidase & mAb3.

The second example shown (Fig. 3B: glucose oxidase, mAb3, and mAb1) was more challenging in terms of the similarity of spectra of the two antibodies used. When linearly correlated, their spectra had an R^2 of 0.998. The solid lines in Figure 3B (right) show this similarity as both spectra are almost indistinguishable visually. Nevertheless, the number of components was reliably found to be three and retention times had an accuracy of 2.1, 0.4, and 1.0 s. RMSE values for the extracted spectra were 0.025, 0.018, and 0.008.

The third example (Fig. 3C: mAb4, glucose oxidase, & mAb3) was among the most challenging 3-component systems. The retention time difference between first and last component was only 10.1 s and the spectra were rather similar ($R^2 = 0.976$ for mAb4 & glucose oxidase and $R^2 = 0.964$ for glucose oxidase and mAb3). For that reason, determined retention times only had an accuracy of 2.8, 1.8, and 1.4 s.

The average accuracy of calculated retention times over all 13 3-component systems was 0.66 s, pure component spectra had an average RMSE of 0.011, both values slightly increased compared to those for the 2-component systems.

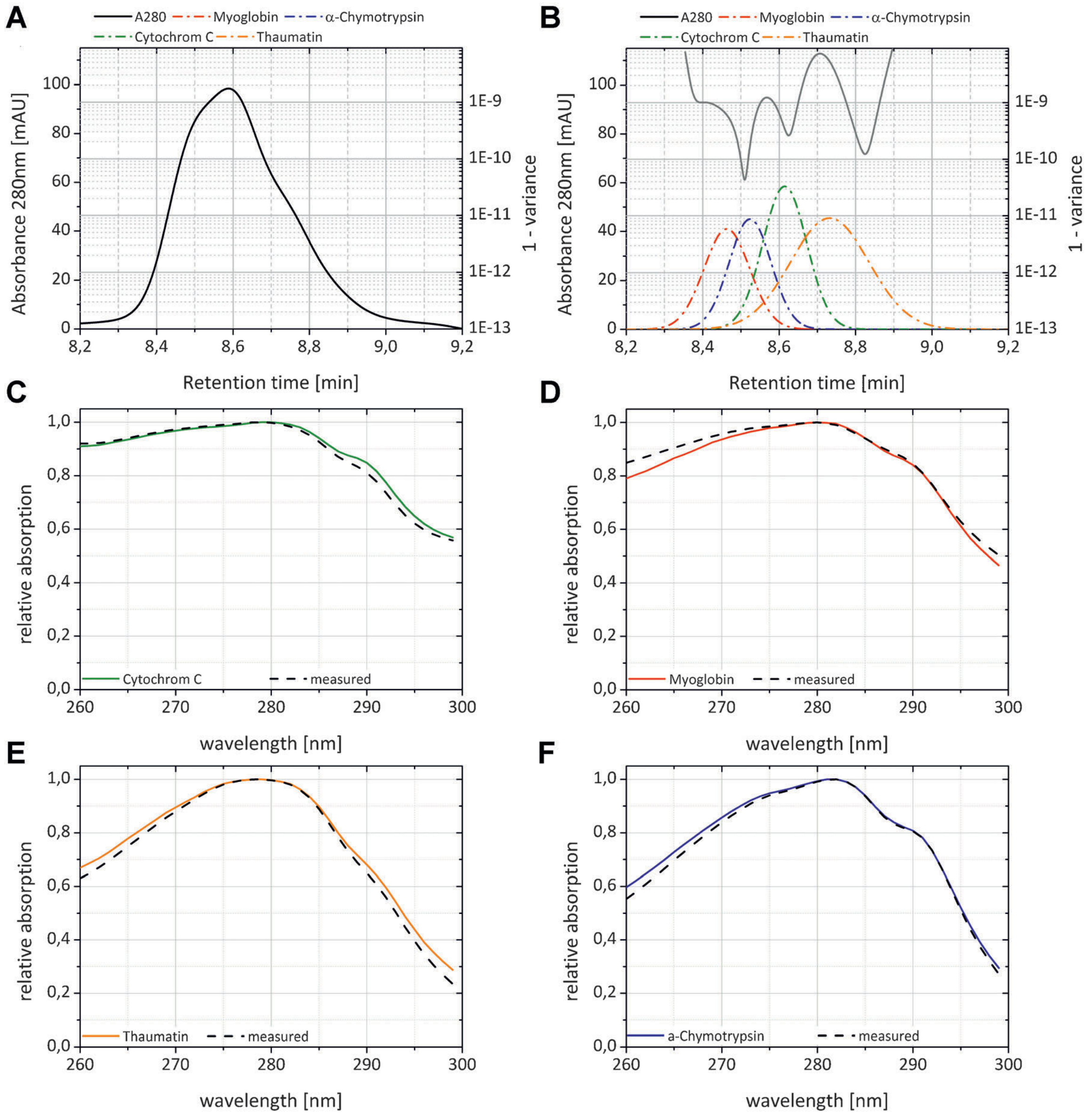

Figure 4. 4-component system of myoglobin, α-chymotrypsin, cytochtrome *c* & thaumatin. **A**: UV absorption trace at 280 nm. **B**: Elution peaks of single components and 1-variance of first principal component. **C–F**: Pure component spectra for all four components from single injections (solid lines) and approximated (dashed lines).

4-Component Systems

A total of seven systems consisting of four co-eluting components were the most complex systems studied. Retention time differences between first and last peak were in the range of 8.4–22.7 s with an average peak resolution of 0.16. Retention times could be determined with an average accuracy of 0.58 s and an average RMSE of 0.014.

One representative example is shown in Figure 4 consisting of myoglobin, α-lactalbumin, cytochrome *c*, and thaumatin. Peak retention times had an accuracy of 1.4, 0.2, 0.1, and 0.2 s and pure component spectra could be approximated with RMSE values of 0.007, 0.008, 0.011, and 0.008.

Sources of Error

Several parameters have an effect on the quality of retention time and peak area determination as well as pure component spectra approximation. To study this systematically, we used three different 2-component systems that differ in the similarity of pure-component spectra (R^2 of 0.72 for HSA & avidin, 0.84 for α-lactalbumin & cytochrome *c*, and 0.997 for glucose oxidase and catalase) (Fig. 5). Since the data were

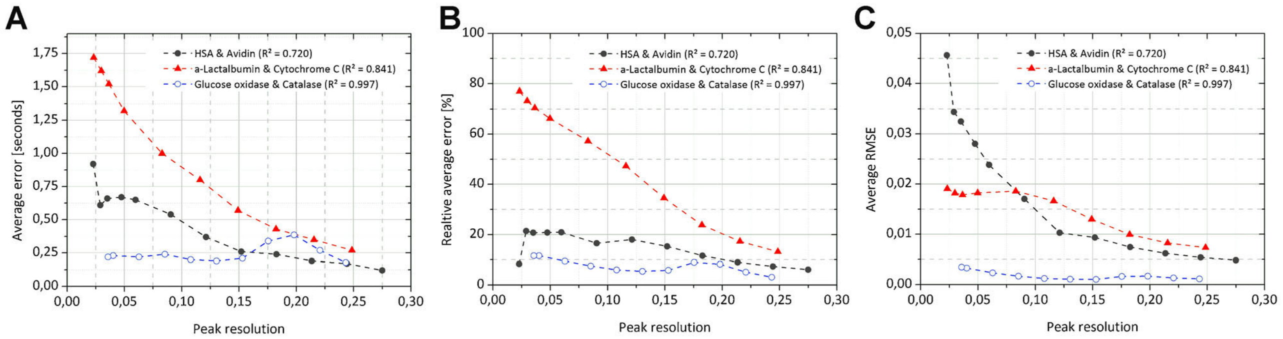

Figure 5. Systematic error evaluation using three different 2-component systems that differ in the similarity of pure component spectra (indicated by R^2 values). **A**: Average error in retention time (**B**) Average error of peak area. **C**: Average RMSE value of approximated and real pure-component spectra.

generated from single component runs, we were able to generate chromatograms with different peak resolutions simply by shifting the resolution peak of one of the two components. This was done in order to investigate the accuracy of the method with increasing co-elution.

Figure 3 shows the effects of peak resolution on retention time (A), peak area (B), and pure component spectra determination (C). In all three systems, two components were reliably found for peak resolutions down to 0.023 (about 0.7 s retention time difference) for the first two systems and 0.036 (1.6 s) for the third system. As expected, the error increased when peak resolution decreased. Interestingly, error values did not increase systematically with decreasing spectral similarity as the α-lactalbumin/cytochrome *c* system showed the highest error while the most challenging system (glucose oxidase & catalase) had the lowest error. Peak area determination (Fig. 3B) showed similar trends.

One possible explanation is, that if pure component spectra are very different, an accurate first approximation of pure-component spectra is essential for the initial concentration profiles of both components, thus small inaccuracies might have a more pronounced negative impact on the accuracy of the results. If pure component spectra are similar, initial guesses are often closer to the real spectra, as can be seen in Figure 3C: the average RMSE for glucose oxidase & catalase is comparably low and essentially independent of peak resolution as each possible sum spectrum of both proteins is already very similar to the pure component spectra.

An additional source of error is the peak area ratio of both components. To study this systematically we changed the relative amount of human lysozyme in a lysozyme/subtilisin co-elution up to a peak area ration of 45:1. For higher ratios, only one component could be identified in the elution peak. For all ratios where both components were identified, the average error in retention time increased only from 0.2 to 1.4 s. While the peak area determination improved for lysozyme from 12% to 3%, it increased for subtilisin from 5% to about 180% for the extreme case of a 45:1 mixture.

Peak Area Determination

Compared to the accurate determination of retention times, peak area calculation showed relatively high deviations. For the 2-component systems peak areas could be determined with an average error ±15.7% that increased to ±29.9% for the 3-component systems. For the 4-component systems the error was ±23.8%. For an accurate peak area determination, the initial approximation of pure component spectra is crucial as these spectra are used to calculate the initial concentration profiles for all individual components. Inaccurate approximation of the spectra leads to rather broad concentration profiles and thus broad Gaussian peaks after fitting. As discussed above, the accuracy of peak area determination is also a function of peak area ratio of the co-eluting components. As a consequence, the main potential of our method lies in the accurate determination of retention times, for example, crucial in the determination of parameters for modeling chromatography, where absolute quantitative data in not needed. Purity analysis is limited to the determination of the number of contaminants rather than their accurate amount.

Application: SEC to SDS

As described in the Introduction Section, one possible application of the spectral method lies in the field of parameter determination of chromatographic model parameters. In an approach by Ahamed et al. (2007), 2D chromatography was used for parameter determination of a complex protein mixture. The first dimension was pH gradient elution to reduce sample complexity, the second dimension was ion-exchange chromatography with multiple gradient slopes for parameter estimation. Since in the second dimension proteins often co-elute, SDS page gel electrophoresis was used for fractions from the second dimension to identify components. Standard SDS page gel electrophoresis is both time-consuming and can strongly depend on the sample buffer conditions, especially the salt concentration. Thus, fast high resolution SEC would be preferable.

To demonstrate the potential of our method, we generated virtual SDS gels that are shown in Figure 6. Figure 6A shows an SEC chromatogram of 18 out of 22 proteins used in this study. Since our spectral method was only validated for four co-eluting compounds, we could not include all 22 proteins (mAb1, mAb3, thaumatin, α-lactalbumin were not included). By visual inspection, only seven distinct maxima can be found when a mixtures of all component is injected. Figure 6B shows three virtual electrophoresis gels: the one on the left is an ideal gel generated with data from single injections on the SEC column, bands indicate peak maxima. The gel in the middle was generated from the A_{280} trace that showed only seven distinct maxima. The gel on the right was generated using our spectral analysis for retention time determination. All 18 components were successfully identified and retention times could be determined. This shows the applicability of an SEC in the second dimension for the identification of coeluting components, for example, for SMA parameter determination in a complex mixturem. In this case, retention time determination is sufficient, correct concentration profiles and peak areas are not needed.

Conclusions

In the work presented, we could show that spectral analysis of co-eluting compounds can significantly increase the analytical capabilities of chromatography even for biological compounds where the components to be analyzed often share a high degree of spectral similarity. The number of components could reliably be determined independent of the number of components (co-elution of up to four components were investigated). Retention times in 2-, 3-, and 4-component systems could be approximated with an average error of about 0.60, 0.66, and 0.58 s. No systematic trend of an increasing error with increasing number of components could be observed. Additionally, pure component spectra could be approximated from overlapping peaks with an average RMSE of 0.011 for the 2-component systems, increasing with the number of proteins up to 0.014.

Possible sources of error were systematically studied for three 2-component systems. As expected the error increased with decreasing peak resolution. Nevertheless, retention times could be calculated for proteins eluting within 0.75 s if pure component spectra were not too similar ($R^2 < 0.85$). Peak area ratios were also identified to influence accuracy. The effect on retentions time determination was not very pronounced (average error increased to 1.4 s for a peak area ratio of 45:1). The effect on peak area determination was more pronounced, especially for the low concentrated component. Overall, the accuracy of peak area determination was in the range of 15.7–29.9% depending on system complexity.

The spectral method presented here is, in principle, applicable to all modes of chromatography, although data shown was only for size exclusion chromatography. Even UV absorbing components in a linear gradient should not pose a major problem, since the method presented here is based on discontinuous spectral changes over time. Nevertheless, sensitivity is supposed decreased. This will be part of future investigations.

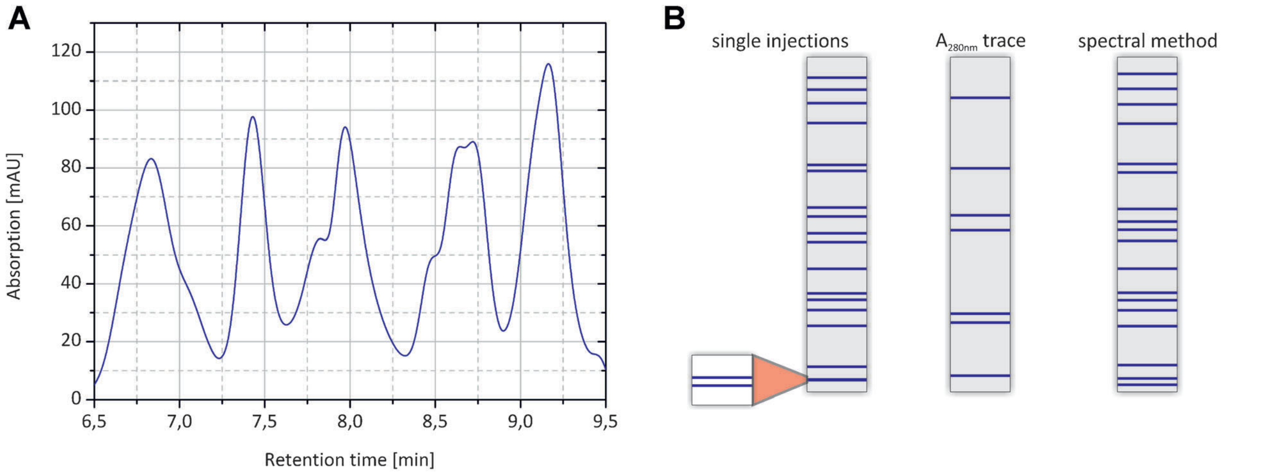

Figure 6. **A**: SEC chromatogram of a mixture of 18 proteins with only seven distinct maxima. **B**: Virtual SDS gels, (**left**) based on single injections; (**middle**) based on the UV trace and (**right**) based on our spectral method.

References

Ahamed Tangir, Nfor BK, van de Sandt Emile JAX, Eppink Michel HM, Verhaert Peter DEM, van Dedem Gijs WK, van der Wielen Luuk AM, Ottens M. 2007. BIOT 469-fast acquisition of bioseparation process development data from crude protein mixtures. Abstr Pap Am Chem S 234.

Ahamed T, Nfor BK, van der Wielen LAM, Verhaert PDEM, van Dedem GWK, Eppink MHM, van de Sandt EJAX, Ottens M. 2009. Omics tools in accelerated purification process development: Multidimensional fractionation and characterization of crude protein mixtures. J Biosci Bioeng 108:S58.

Bailey HP, Rutan SC. 2011. Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine. Chemom Intell Lab Syst 106:131–141.

Berthod A. 1991. Mathematical series for signal modeling using exponentially modified functions. Anal Chem 63:1879–1884.

Brooks CA, Cramer SM. 1992. Steric mass-action ion exchange: Displacement profiles and induced salt gradients. AIChE J 38:1969–1978.

Bu D, Brown CW. 2000. Self-modeling mixture analysis by interactive principal component analysis. Appl Spectrosc 54:1214–1221.

Caballero RD, García-Alvarez-Coque MC, Baeza-Baeza JJ. 2002. Parabolic-Lorentzian modified Gaussian model for describing and deconvolving chromatographic peaks. J Chromotogr A 954:59–76.

Coffman JL, Kramarczyk JF, Kelley BD. 2008. High-throughput screening of chromatographic separations: I. Method development and column modeling. Biotechnol Bioeng 100:605–618.

Foley JP, Dorsey JG. 1983. Equations for calculation of chromatographic figures of merit for ideal and skewed peaks. Anal Chem 55:730–737.

Giddings JC. 1967. Maximum number of components resolvable by gel filtration and other elution chromatographic methods. Anal Chem 39:1027–1028.

Guillarme D, Ruta J, Rudaz S, Veuthey J-L. 2010. New trends in fast and high-resolution liquid chromatography: A critical comparison of existing approaches. Anal Bioanal Chem 397:1069–1082.

Hanggi D, Carr PW. 1985. Errors in exponentially modified Gaussian equations in the literature. Anal Chem 57:2394–2395.

Hansen SK, Skibsted E, Staby A, Hubbuch J. 2011. A label-free methodology for selective protein quantification by means of absorption measurements. Biotechnol Bioeng 108:2661–2669.

Lan K, Jorgenson JW. 2001. A hybrid of exponential and gaussian functions as a simple model of asymmetric chromatographic peaks. J Chromotogr A 915:1–13.

Lawson CL, Hanson RJ. 1974. Chapter 23. In: Solving Least Squares Problems. Upper Saddle River, New Jersey: Prentice-Hall. p 161.

Lawton WH, Sylvestre EA. 1971. Self modeling curve resolution. Technometrics 13:617.

Liapis AI. 1990. Modelling affinity chromatography. Sep Purif Rev 19:133–210.

Naish PJ, Hartwell S. 1988. Exponentially modified Gaussian functions—A good model for chromatographic peaks in isocratic HPLC? Chromatographia 26:285–296.

Osberghaus A, Hepbildikler S, Nath S, Haindl M, von Lieres E, Hubbuch J. 2012a. Optimizing a chromatographic three component separation: A comparison of mechanistic and empiric modeling approaches. J Chromotogr A 1237:86–95.

Osberghaus A, Hepbildikler S, Nath S, Haindl M, von Lieres E, Hubbuch J. 2012b. Determination of parameters for the steric mass action model—A comparison between two approaches. J Chromotogr A 1233:54–65.

Osberghaus A, Drechsel K, Hansen S, Hepbildikler SK, Nath S, Haindl M, von Lieres E, Hubbuch J. 2012c. Model-integrated process development demonstrated on the optimization of a robotic cation exchange step. Chem Eng Sci 76:129–139.

Osten DW, Kowalski BR. 1984. Multivariate curve resolution in liquid chromatography. Anal Chem 56:991–995.

Piaggio MV, Peirotti MB, Deiber Ja. 2005. Effect of background electrolyte on the estimation of protein hydrodynamic radius and net charge through capillary zone electrophoresis. Electrophoresis 26:3232–3246.

Tejeda-Mansir A, Montesinos RM, Guzmán R. 2001. Mathematical analysis of frontal affinity chromatography in particle and membrane configurations. J Biochem Biophys Methods 49:1–28.

Vandeginste B, Essers R, Bosman T, Reijnen J, Kateman G. 1985. Three-component curve resolution in liquid chromatography with multiwavelength diode array detection. Anal Chem 57:971–985.

Wright PR, Muzzio FJ, Glasser BJ. 1998. Batch uptake of lysozyme: Effect of solution viscosity and mass transfer on adsorption. Biotechnol Prog 14:913–921.