**Real-time prediction of protein quantity and purity in downstream processing**

Michael Melcher1,2,\*, Theresa Scharl-Hirsch1,2, Dominik Sauer1,3, Nicole Walch1,3, Edit Felföldi1,3, Anna Christler1,3, Alois Jungbauer1,3, Astrid Dürauer1,3 and Friedrich Leisch1,2

*1 Austrian Centre of Industrial Biotechnology (ACIB), Muthgasse 11, 1190 Vienna, Austria; 2 Institute of Applied Statistics and Computing, University of Natural Resources and Life Sciences Vienna, Peter-Jordan-Straße 82, 1190 Vienna, Austria; 3 Department of Biotechnology, University of Natural Resources and Life Sciences Vienna, Muthgasse 18, 1190 Vienna, Austria*

*\*Corresponding author: michael.melcher@acib.at*

**Highlights**

* Online-monitoring of capture step of human fibroblast growth factor (FGF) 2
* Multi-analytical approach (Fluorescence, ATR/FTIR, UV, refractive index, …)
* Various statistical learning methods (PLS, Lasso, structured additive regression)
* Prediction errors of approx. 0.5 mg/ml (protein quantity), 200 ppm (host cell protein) and 300 ppm (dsDNA)

**1. Introduction**

Downstream processing (DSP) describes the purification of biopharmaceuticals with a chromatographic step being the main part. Conventional online sensors such as UV/Vis, pH or conductivity probes are currently used for monitoring this process. Unfortunately, key process parameters such as the concentration of various impurities (e.g. host cell protein HCP or double-stranded DNA, dsDNA) are thereby not accessible requiring time-, labor- and cost-intensive offline measurements of the fractionated eluate [1].

**2. Methods**

A conventional chromatographic workstation (Äkta Pure 25, GE Healthcare, Sweden) equipped with a multi-wavelength UV/Vis detector (the wavelengths 214, 260 and 280 nm are monitored), a conductivity and a pH probe, was enhanced with a mid-IR spectrometer (MATRIX-FM, Bruker, USA), a fluorescence detector (AvaSpec-ULS-TEC, Avantes, Netherlands), a multi-angle light scattering (MALS) spectrometer (miniDAWN TREOS, Wyatt, USA) as well as with a differential refractive index (RI) detector (Optilab T-rEX, Wyatt, USA). The chromatographic capture step of human fibroblast growth factor 2 (FGF-2) expressed in *E. coli* was monitored. On- and corresponding offline data for protein concentration and purity (HCP and dsDNA) were measured for 16 runs resulting in an extensive data set. Traditional chemometric methods such as partial-least squares (PLS) or principal component regression (PCR) were compared with penalized techniques (ridge and Lasso regression) and structured additive regression (STAR) [2,3]. All computations were performed in the R statistical computing environment [4].

**3. Results and discussion**

Figure 1 depicts the performance of a quantity model on an independent test run showing good agreement between model predictions and measured values determined for the 15 fractions. On average (over 6 test runs) an accuracy of 0.51 mg/ml (root-mean-squared error) is observed and can be expected for future applications of the model. A reduction of more than 30% in prediction error can be achieved compared by using spectroscopic data compared to models based on UV, conductivity and pH as predictors. Similarly, for HCP and dsDNA errors of approximately 200 and 300 ppm, respectively, are obtained.

F:\Projekt_25011\ECAB5\plot_ecab.tiff

**Figure 1.** Performance of a structured additive regression (STAR) model for the protein concentration on a new test run – online prediction on a time grid of 1 second (black) and concentration values determined by offline measurements for 15 fractions of 1 ml each (red). Five minutes before and after the elution peak (15 min) are shown.

**4. Conclusions**

Particularly non-linear STAR models in combination with boosting as a variable selection method allow for a reliable prediction of critical process quantities such as the protein concentration, host cell protein and dsDNA content. While traditional sensors (UV/Vis, pH and conductivity) might be sufficient for the protein concentration, spectroscopic sensor systems are essential for modeling the impurities.

**References**

1. D. G. Sauer, M. Mosor, A.-C. Frank, F. Weiß, A. Christler, N. Walch, A. Jungbauer, A. Dürauer, Protein Expr. Purif. 153 (2019) 70-82.
2. R. Tibshirani, J. Royal Stat. Soc. B 58 (1996) 267-288.
3. M. Melcher, T. Scharl, M. Luchner, G. Striedner, F. Leisch, Biotechn. Bioeng. 114 (2017) 321-334.
4. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>, 2018.