**Using Large Data for the Prediction of Quality Attributes of an Antibody Capture Process in Real-Time**

Theresa Scharl-Hirsch1,2,\*, Michael Melcher1,2, Edit Felföldi1,3, Dominik Sauer1,3, Nicole Walch1,3, 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: theresa.scharl@acib.at*

**Highlights**

* Online monitoring in downstream processing.
* Real-time prediction of critical quality attributes.
* Machine learning application on large data from biotechnology.
* Pooling algorithm for material collection of an antibody capture process.

**1. Introduction**

It is state of the art in biopharmaceutical industry to monitor chromatographic processes by standard detectors and probes such as UV/VIS, pH and conductivity. Quality attributes such as product quantity and purity (e.g. dsDNA content, host cell protein concentration and high molecular mass impurities) can only be obtained by collecting and analysing fractions after each unit operation. In our project we have developed an online monitoring system for downstream processes, which is based on an array of online sensors. ATR-FTIR spectrometer, a fluorescence detector as well as a multi angle light scattering detector and a refractive index detector were implemented as additional sensors.

**2. Methods**

In order to use all available online signals in real time a powerful database environment is needed. Here we use an evon database (<https://en.evon-automation.com/xamcontrol/>) to store all the data that is measured by the sensors. A quantitative protein A run takes several hours. For the implemented sensors more than 18.000 online signals are measured in a time grid of one second yielding a huge amount of data. For further data usage efficient storage, time alignment and data pre-processing are crucial. In a first step several antibody capture runs were conducted where online signals of all available sensors were measured during the whole run. In addition, fractions were collected where quality attributes were analysed. In the next step all available online and offline data were processed using the statistical computing environment R [1].

**3. Results and discussion**

Several machine learning methods (e.g., partial least squares, random forests, generalised additive models) were used to find a relation between the measured offline variables such as product concentration or dsDNA content and the available online variables. Structured additive regression models in combination with boosting as a variable selection tool were found to be a useful modelling technique [2]. These prediction models allow online pooling decisions replacing time and labor-intensive laboratory measurements (Figure 1).



**Figure 1.** Measured (A) and predicted (B) quality attributes of an antibody capture run.

**4. Conclusions**

Prediction models solely based on online signals were set up providing real-time predictions. These models can be directly applied in real-time inside the database during a new chromatographic run to predict the quality attributes of a product solely based on the online variables measured. From the quality perspective it is also in accordance with FDA’s recommendation. It is requested to include on-line sensors to improve the constant quality of the product.

**References**

1. 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.
2. Melcher, M., Scharl, T., Luchner, M., Striedner, G., Leisch, F., Boosted structured additive regression for Escherichia coli fed-batch fermentation modeling. Biotechnology and Bioengineering, 2017, 114, 321-334.