Is SVM a black box? Visualization and interpretation of SVR models to analyze glycosylation in pharmaceutical proteins (mAbs, Fc) based on infrared spectroscopy FT-IR

Sabrina Hamla[1]*, Pierre-Yves Sacré[1], Allison Derenne[2], Kheiro-Mouna Derfoufi[2], Ben Cowper[3], Claire I. Butré[4], Arnaud Delobel[4], Erik Goormaghtigh[2], Philippe Hubert[1], Eric Ziemons[1].

[1] University of Liege (ULiege), CIRM, Vibra-Sante Hub, Department of Pharmacy, Laboratory of Pharmaceutical Analytical Chemistry, Liege, Belgium.

[2] Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes, ULB, Campus Plaine CP206/02, 1050 Brussels, Belgium.

[3] National Institute for Biological Standards and Control, Blanche Lane, South Mimms, Potters Bar, Hertfordshire, EN6 3QG, United Kingdom.

[4] Quality Assistance, Techno Parc de Thudinie 2, 6536 Thuin, Belgium.

Keywords: Fourier Transform Infrared Spectroscopy (FT-IR), Pharmaceutical proteins, Monosaccharides content, Kernel functions, P-vector, Non-linear Support Vector Regression (SVR), Partial least squares (PLSR).

1 Introduction

Almost 60% of commercial pharmaceutical proteins are glycosylated. Glycosylation is considered a critical quality attribute as it affects proteins' stability, bioactivity, and safety. Hence, the regulatory authorities require systematic characterization of the composition and structure of glycoproteins throughout development processes. Currently existing methods are time-consuming, expensive and require significant sample preparation steps which can alter the robustness of the analyses. Thus, we have suggested the use of a fast, direct and simple method based on Fourier transform infrared spectroscopy (FT-IR) and Support Vector Regression (SVR) [1]. The SVR method is considered a powerful alternative to Partial least squares (PLSR), thanks to its ability to model nonlinear relationships flexibly. Usually, SVR is applied as a black-box method, unlike the PLSR where the models can be straightforwardly interpreted by evaluating their loading. Thus, the aim of the present study was to succeed in visualizing and interpreting the SVR models to quantify glycosylation by adapting the methodology proposed by Üstün et al. on the radial basis function (RBF) kernel [2].

2 Theory

Consider an input data set $X(N \times M)$ with an output vector $y_i \in R$. The objective of SVR is to find a multivariate regression function $f(x)$ to predict the desired output property (amount of monosaccharides) of an unknown object (spectrum).

$$f(x) = \sum_{i,j=1}^{N} (\alpha_i - \alpha_i^*)\, K(x_i, x_j) + b \qquad (1)$$

α_i and α_i^* correspond to the support vectors ($\alpha_i, \alpha_i^* \neq 0$). $K(x_i, x_j)$ is a RBF kernel that transforms the non-linear input space into a high-dimensional feature space, where the problem can be modeled in a linear way. However, the information related to the original input variables is lost. This is why, the first step was to calculate the correlation matrix $R(N \times M)$ between each column of the input data (spectral variables) and each row of the kernel matrix (similarity measure), in order to clearly determine which input variables in the original input data are explanatory for the modeled output property (amount of monosaccharides) (Fig. 1.A). The second step was to determine the contribution of each input variable to the final regression model, a P-vector is obtained by calculating the inner-product between the original input space (XT) and the α-vector of the SVR models (Fig. 1.B) [2].

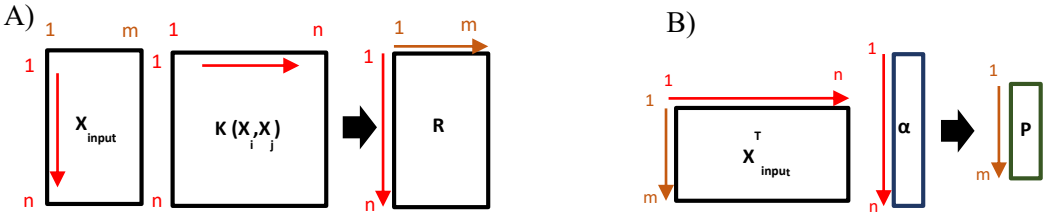

Figure 1 – A) illustrates the correlation matrix $R(N \times M)$ and B) illustrates the P-vector.

3 Material and methods

Thirty-two therapeutic proteins were investigated. Micro-Biospin™ columns were used to remove the excipients which can interfere with the FT-IR measurements. Samples were analyzed by Bruker Tensor 27 FT-IR spectrophotometer (Bruker Optics GmbH, Ettlingen, Germany) with the software Opus 6.5 (Bruker Optics GmbH, Ettlingen, Germany). 0.5 µL of each protein sample was deposited and dried on the diamond crystal of the ATR device. For each glycoprotein, six spectra for each three independent samples were recorded between 4000 and 600 cm^{-1} at a resolution of 2 cm^{-1} and 128 scans.

The N-glycans of the antibodies were characterized by GlycoWorks RapiFluor-MS™ kit (Waters, MA, USA) and reference data were performed by UPLC-FLR-MS analysis. Regression models (PLSR, SVR) and P-vector were computed on mixtures of monosaccharides and samples of glycoproteins. Matlab® (R2017b) and the PLS-Toolbox (Eigenvector Research, WA, USA) were used. All the spectra were preprocessed by Savitzky-Golay 2nd derivative (polynomial order: 3, window size: 15) followed by standard normal variate (SNV).

4 Results and discussion

For example, in the mannose mixture model, mannose characteristic bands were present in LV1 (PLS) and also in P-vector (build-in SVR). The same results were observed for the glycoproteins model.

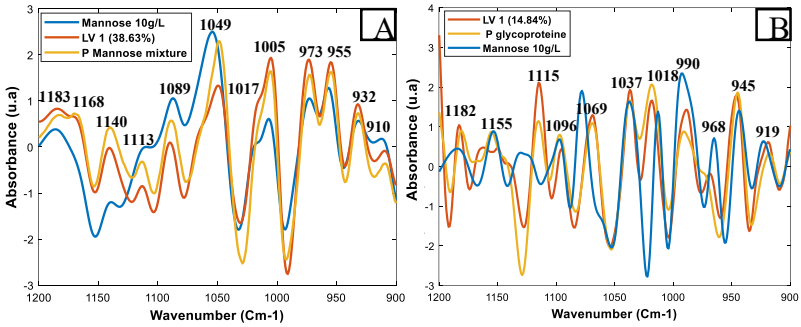

Figure 2- illustrates the overlay of the spectra of pure mannose, LV1, P-vector. A) in a mixture of mannose and B) in the samples of glycoproteins.

5 Conclusion

It can be concluded that this visualization and interpretation approach facilitates the understanding and specificity analysis of SVR models to quantify monosaccharides content in glycoproteins.

6 References

[1] A. Derenne. FTIR spectroscopy as an analytical tool to compare glycosylation in therapeutic monoclonal antibodies. Anal. Chim. Acta. 2020, pp. 62–71.

[2] B. Üstün. Visualisation and interpretation of Support Vector Regression models. Anal. Chim. Acta. 2007, pp. 299–309.