Analytica Chimica Acta 825 (2014) 42–50

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Raman model development for the protein conformational state classification in different freeze-dried formulations

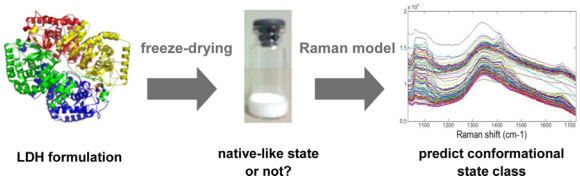
Sigrid Pieters [a], Jean-Michel Roger [b], Thomas De Beer [c], Matthias D'Hondt [d], Bart De Spiegeleer [d], Yvan Vander Heyden [a,*]

[a] Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research, Vrije Universiteit Brussel – VUB, Laarbeeklaan 103, Brussels B-1090, Belgium
[b] Irstea, UMR1201 ITAP, Rue J.F. Breton 361, Montpellier 34191, France
[c] Laboratory of Pharmaceutical Process Analytical Technology, Ghent University, Harelbekestraat 72, Ghent B-9000, Belgium
[d] Laboratory of Drug Quality & Registration (DruQuaR), Ghent University, Harelbekestraat 72, Ghent B-9000, Belgium

HIGHLIGHTS

- Raman model for predicting protein conformational state class was developed.
- Formulation and batch variability on the Raman spectra were investigated.
- Knowledge about spectral variability origin necessary for robust model development.
- External parameter orthogonalization is a valid alternative to exhaustive calibration.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 January 2014
Received in revised form 18 March 2014
Accepted 21 March 2014
Available online 26 March 2014

Keywords:
Raman spectroscopy
Multivariate classification
Robustness
External parameter orthogonalization
Formulation
Batch

ABSTRACT

The aim of this work is to build a multivariate calibration (MVC) model from Raman spectra for the prediction of the protein conformational state class (i.e. native-like or non-native) in different freeze-dried pharmaceutical formulations of a model protein lactate dehydrogenase (LDH). As this model would be intended to facilitate and better understand formulation and process development, it should allow acceptable classification performance despite variations in formulation type and batch. Therefore, it was attempted to (1) find which factors interfere the Raman spectra, (2) understand them, and (3) make the MVC model robust for them. A variance analysis within the Raman spectral data space identified significant spectral background variations among certain formulation types and batches in the studied samples. Raw material (i.e. LDH) batch variability and the presence of a Maillard reaction in formulations were the main reasons for this. We demonstrate the successful use of both exhaustive calibration and external parameter orthogonalization (EPO) pre-processing for making the Raman classification model more robust for the expected spectral interferences.

1. Introduction

Multivariate calibration (MVC) models to predict quality attributes from spectroscopic (e.g. Near-infrared, Raman) measurements of samples are gaining popularity in the (bio)

* Corresponding author. Tel.: +32 2 477 47 34; fax: +32 2 477 47 35.
E-mail address: yvanvdh@vub.ac.be (Y.V. Heyden).

http://dx.doi.org/10.1016/j.aca.2014.03.027
0003-2670/

pharmaceutical field [1,2]. Their fast and simple analyses are main contributors to their success. Because of its high sensitivity for protein conformation, Raman spectroscopy combined with multivariate analysis (MVA) has been proposed for the rapid assessment of the conformational state in freeze-dried proteins [3]. Such an MVC model can potentially be useful for speeding up freeze-drying formulation and process development provided that it enables accurate classification in different formulations and batches. The latter will be studied in this paper.

The typical procedure for building an MVC model is obtaining a number of representative calibration samples and collecting a spectrum for each of them as well as determining the quality attribute of interest through a reference method. It is then aimed to describe the quality attribute ($\boldsymbol{y}$), expressed as a numerical value or as a dummy variable related to its class membership, as a function of the measured spectral signals ($\boldsymbol{X}$). The partial least squares (PLS) algorithm that captures efficiently the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$ has become the most popular means for developing multivariate regression models [4]. It is also frequently used as a dimensionality reduction technique in conjunction with classification algorithms, such as discriminant analysis (DA) [5]. In this paper a PLS-LDA classification is considered.

In most cases, only a fraction of the spectral variation in $\boldsymbol{X}$ is correlated to $\boldsymbol{y}$. This is the predictive (or correlated) spectral variation. The net analyte signal (NAS) is defined as the useful part of the raw signal for the prediction of the quality attribute of interest. Therefore it theoretically corresponds to 100% predictive variation, while being orthogonal to the variation of all other spectral contributors [6,7]. The other variations within the $\boldsymbol{X}$ variable space constitute of orthogonal (or uncorrelated) spectral variability [8–10]. This variability may find its origin in all other spectrally contributing factors, including spectral noise and systematic variations caused by external factors. The complexity of this orthogonal variation will largely depend on the type of spectroscopic technique and the application. For instance, spectral variations can be the result of varying physical, chemical, instrumental or process factors. One example is spectral background variation. In Raman spectra this may arise from non-Raman effects (such as fluorescence) [11,12], while in diffuse reflectance NIR spectra light scattering effects of particles and a variable path length may produce a spectral background.

As PLS is an efficient tool to recognize most of the predictive variation, it may enable selective and accurate analysis in the presence of strong interferences, such as in complex matrices or processes [13]. However, PLS-based calibration models will pick up any correlation that can be found between $\boldsymbol{X}$ and $\boldsymbol{y}$, regardless the origin of the spectral changes. As a result, conditions changing the orthogonal spectral variation, can make the calibration model lose its prediction accuracy for the quality attribute when that type of orthogonal variation was not adequately considered in the calibration set [14]. This necessitates a calibration set being representative for future specimens and therefore also requires the identification of factors contributing to the orthogonal systematic spectral variation. Calibration samples should be selected in such a way that not only the variation related to the quality attribute of interest, but also the expected orthogonal variation, varies over a range that is expected to be present in future specimens (i.e. exhaustive calibration) [15]. Composing a representative calibration set can therefore become a challenging and costly task, especially if there are many potential sources of spectral variability. Pre-processing of the spectra in $\boldsymbol{X}$ is another way to diminish the effects of known external factors hampering the model robustness. For instance, background corrections (e.g. (linear) interpolation baseline correction methods, MSC, SNV . . .) have been applied to remove the background contributions of the experimentally obtained Raman spectra prior to interpretation or modeling [11,12]. Another alternative for making the calibration set representative is using the calibration base after filtering the orthogonal space containing the identified harmful variability. External parameter orthogonalization (EPO) has been used for this purpose in order to make NIR models more robust [9,16–18].

Robust MVC model development requires thus the knowledge of the external factors that systematically may influence the spectral variability [14,19]. In the present study, the first aim was to identify and understand the factors influencing significantly the class-orthogonal spectral variability. 'Formulation type' was chosen as the first factor because the model should be able to predict the conformational state of the proteins in various model formulations. 'Batch' was the second factor because the MVC model should allow the correct classification of samples from new production batches. In second part of this paper the robustness of a PLS-LDA model for these factors was evaluated with external test sets. Both exhaustive calibration and different pre-processing methods (with no user intervention – i.e. no baseline subtraction methods – to exclude a user-dependent bias) were evaluated for making the PLS-LDA model more robust when predicting the protein conformational state class in different formulations and batches.

2. Theory

2.1. External parameter orthogonalization (EPO)

Considering the contributions within the calibration matrix $\boldsymbol{X}$, i.e. predictive variation for the analyte k ($\boldsymbol{X}_k$), orthogonal variation from all other sources ($\boldsymbol{X}_{-k}$), and the random spectral variation ($\boldsymbol{E}$), Eq. (1) can be written.

$$X = X_k + X_{-k} + E \tag{1}$$

EPO uses the spectra from a small experimental design to define a basis of the space spanned by the 'interfering' external factor(s), this way estimating the parasitic subspace $\boldsymbol{X}_{-k}$. Hereby, the external factor is varying while the quality attribute of interest stays constant. In other words, for a set of n samples, n spectra are acquired at p levels of the external factor. Mean centering each set of p spectra removes the information of the quality attribute of interest, hence only the spectral variations due to the external factor remain. Then the matrix $\boldsymbol{D}$ is composed by merging the p mean-centered spectra from each of the n samples. Principal component analysis (PCA) is then applied to $\boldsymbol{D}$. Retaining only the first g principal components (PCs), the column vectors of the matrix of eigenvectors $\boldsymbol{G}$ will represent an orthonormal basis of the subspace to be removed. Finally, an orthogonal projection is defined to filter the calibration spectra $\boldsymbol{X}$ in order to obtain the 'corrected' ones ($\boldsymbol{X}^*$).

$$X* = X(I - GG^T) \tag{2}$$

where $\boldsymbol{G}$ is a matrix comprising the g first eigenvectors of the square matrix $\left[D^T D\right]$, and $\boldsymbol{I}$ is the identity matrix [16].

3. Methods and materials

3.1. Materials and equipment

Two batches of L-lactic dehydrogenase (LDH) from rabbit muscle – Type II in 'ammonium sulfate were obtained from Sigma–Aldrich (Saint Louis, MO, USA). Due to a shortage of the type of LDH used, it was not possible to obtain more LDH raw material batches from the supplier. From this raw material, different LDH formulation types were prepared (Table 1). Prior to freeze-drying, 2 mL type I vials (Nipro, Authon-du-Perche, France) were filled