Talanta 107 (2013) 368–375

Contents lists available at SciVerse ScienceDirect

Talanta

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

Modified locally weighted—Partial least squares regression improving clinical predictions from infrared spectra of human serum samples

David Perez-Guaita [a], Julia Kuligowski [b], Guillermo Quintás [c], Salvador Garrigues [a,*], Miguel de la Guardia [a]

[a] Analytical Chemistry Department, University of Valencia, Edifici Jeroni Muñoz, Burjassot, Valencia 46100, Spain
[b] Division of Neonatology, University Hospital Materno-Infantil La Fe, Bulevar Sur, s/n, Valencia, Spain
[c] Leitat Technological Center, Bio In Vitro Division, Terrassa, Spain

ARTICLE INFO

Article history:
Received 23 October 2012
Received in revised form
16 January 2013
Accepted 19 January 2013
Available online 31 January 2013

Keywords:
Local weighted-partial least squares regression (LW-PLSR)
Human serum analysis
Vibrational spectroscopy
Infrared (IR)
Chemometrics

ABSTRACT

Locally weighted partial least squares regression (LW-PLSR) has been applied to the determination of four clinical parameters in human serum samples (total protein, triglyceride, glucose and urea contents) by Fourier transform infrared (FTIR) spectroscopy. Classical LW-PLSR models were constructed using different spectral regions. For the selection of parameters by LW-PLSR modeling, a multiparametric study was carried out employing the minimum root-mean square error of cross validation (RMSCV) as objective function. In order to overcome the effect of strong matrix interferences on the predictive accuracy of LW-PLSR models, this work focuses on sample selection. Accordingly, a novel strategy for the development of local models is proposed. It was based on the use of: (i) principal component analysis (PCA) performed on an analyte specific spectral region for identifying most similar sample spectra and (ii) partial least squares regression (PLSR) constructed using the whole spectrum. Results found by using this strategy were compared to those provided by PLSR using the same spectral intervals as for LW-PLSR. Prediction errors found by both, classical and modified LW-PLSR improved those obtained by PLSR. Hence, both proposed approaches were useful for the determination of analytes present in a complex matrix as in the case of human serum samples.

1. Introduction

Total protein, triglycerides, glucose and urea contents in blood are commonly determined as components of comprehensive metabolic panels typically included in routine health checkups for the evaluation of conditions such as liver and kidney disease or diabetes, among other metabolic and nutritional disorders. On the other hand, the development of multianalyte reagent-free spectroscopy based methods is an active field of research in analytical and clinical chemistry [1,2] where mid-infrared (IR) spectroscopy shows a number of relevant advantages over other techniques based on enzymatic-colorimetric reactions and enzyme-linked immuno sorbent assay (ELISA) determinations. The broad detection capabilities of IR allow its application to the quantification of a big amount of molecules, from small ions to proteins, without previous derivatization steps thus offering a fast, cheap, reagent-free and compactable alternative to enzymatic-colorimetric methods. However, the wide range of molecules with characteristic absorption bands in the mid-IR region increases the likelihood of interferences arising from other sample constituents. Frequently IR absorption bands of matrix components are strongly overlapped with signals of the analytes of interest in the mid-IR region, which hampers the selection of specific bands suitable for analyte quantification based on univariate regressions. In order to develop direct methods, thus avoiding time consuming sample pre-treatment steps, a common approach is the use of multivariate regression [3–9]. In particular, chemometric modeling of spectra provided by attenuated total reflectance (ATR) was applied for the determination of clinical parameters in human serum samples [10].

Multivariate linear regression models, mostly employing partial least squares regression (PLSR), have been developed for the determination of total protein, triglyceride, glucose and urea contents in serum [11–15], plasma [10] and whole blood [16]. A comparison of the analytical performance of these works can be found in [2] and [1].

The aforementioned PLS methods assumed linear relationships between the IR absorbance and the concentration of the analytes within typical physiological ranges. In spite of that, a number of linearity problems associated with ATR measurements in biological samples have been reported in literature [17] due to the formation of biofilms on the ATR cell [18], the confinement of

* Corresponding author. Tel.: +34 96 354 4838; fax: +34 96 354 4845.
E-mail address: salvador.garrigues@uv.es (S. Garrigues).

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.01.035

analytes inside cells [11] or cell sedimentation [19]. Hence, it can be expected that changes in the penetration depth of the IR beam in ATR systems and the complexity of the serum matrix could affect the linearity of the measurements.

PLS is a multivariate inverse method widely used in chemistry [20]. This method extracts a set of latent variables (LVs) explaining the sources of variation in the *X*-block (i.e., spectra) correlated to an *y*-vector (i.e., analyte concentrations). For accurate predictions of analyte concentrations, it is essential to have access to a set of representative calibration samples that include all sources of variation expected to be present in new unknown samples and, at the same time, their spectral contributions to the overall signal should match. Thus, the selection of appropriate calibration sets might be troublesome due to the complex nature of serum samples. Besides, target analyte concentrations typically show a wide variation hindering the availability of appropriate calibration sets.

Addressing the abovementioned concerns, LW-PLSR can be seen as a suitable strategy to overcome a lack of linearity of the relationship between signals and analyte concentrations and to facilitate the selection of proper calibration sets. Local regression approximations are based on the assumption that the use of specific calibration equations for each new sample to be analyzed, using small calibration sets tailored to the unknown sample from a large library of samples, improves prediction accuracy [21]. As described by Pérez-Marin et al. 'local regression, combines the advantages of global calibration in using one database to cover a large product domain, with the accuracy obtainable with specific calibrations' [22]. In other words, for each unknown sample to be predicted, local regression models are based on an initial selection of a reduced set of calibration spectra providing similar features, in order to develop a specific calibration model. Locally weighted regression was first proposed by Cleveland et al. for the estimation of nonlinear regression surfaces when little information about the surface was available [23]. Naes et al. found a nonlinear relationship between sample composition and principal component scores [24,25]. To correct this non-linearity, the same authors applied LW-PLSR, where only the closest samples characterized by a minimum distance in the scores space were employed for local model calculation. Briefly, the approach consisted in four steps: (i) development of a PCA model; (ii) computation of the Euclidean distance between the query and the calibration samples in the scores space, (iii) selection of the nearest neighbors to the query and (iv) calculation of a Principal Component Regression (PCR) or PLSR employing the selected samples. This method has been criticized because it only takes into account "spectral aspects" of the response variable (*X*-block) and ignores the "chemical information" of the concentration vector *y* [26]. Therefore, numerous approaches have been proposed as alternatives for the selection of samples for local models, as for example the similarity estimated from the output of a global method [26] or the Euclidean distance in the *X*-block [27,28]. In addition, methods using the correlation between variables [29] and most recently methods where all samples are stored in the calibration set and weighted depending on their distance to the query [30,31] have been proposed.

The aim of this work is to evaluate the advantages and drawbacks of the application of LW-PLSR to the direct determination of total protein, triglycerides, glucose and urea contents in human serum employing ATR-FTIR spectra. For this propose spectra from 1400 serum samples acquired during a previous study [15] dealing with the determination of the influence of the origin of serum samples on PLS model performance, were used. For the present study, samples from all origins were randomly divided into calibration and validation sets for PLS and LW-PLSR model calculation and comparison of their respective prediction capabilities. In the case of serum samples it must be remarked that the problem of using the PC space to find similarities between samples with similar concentrations of the analyte under study is that results may be influenced by matrix effects. Because of that, a simple approach based on the use of an analyte specific spectral region and a multi-parametric method for the selection of the optimum number of PCs was developed. In addition, a simple, modified LW-PLSR approach is proposed in order to optimize obtained results based on the use of an analyte specific spectral region for PCA clustering to overcome difficulties arising due to matrix effects in complex human serum samples.

2. Materials and methods

2.1. Sample collection and data acquisition

A total of 1400 samples obtained from the Hospital Dr. Peset Alexandre (Valencia, Spain) were analyzed. Reference concentrations of total protein, triglycerides, glucose and urea contents in human serum of the samples included in the study, were obtained through the use of an Abbott Architect c16000 auto-analyzer (Libertyville, IL, USA) as described elsewhere [15] in the clinical laboratory of the hospital, providing precisions $\leq 5\%$. Detailed information about the samples can be found in a previous work [15]. Table 1 summarizes the main descriptive statistical parameters of the sample reference data used throughout this study.

FTIR spectra were acquired on a Bruker Tensor 27 (Bremen, Germany) spectrophotometer equipped with a deuterium triglycine sulfate detector and an ATR DuraSampleIR accessory with a nine reflection diamond/ZnSe Dura disc from Smiths detection Inc. (Warrintong, UK). Aliquots of 150 μL of each serum sample were deposited on the ATR crystal and covered using an N-BK7 PCV lens to avoid sample evaporation. A total of 100 accumulated scans in the 600–4000 cm^{-1} range at a resolution of 4 cm^{-1} were averaged to increase the signal to noise ratio using a background of the empty ATR cell obtained under the same instrumental conditions and ATR correction was applied to the resulting mean spectrum. A blank spectrum of water was subtracted to each serum spectrum. All serum samples were measured by triplicate and averaged. The contribution of water vapor to the final spectrum was subtracted to the average spectrum of each sample. Further details on samples and spectra acquisition are available in [15].

2.2. Software and spectral preprocessing

Data analysis was run under Matlab 7.7.0 from Mathworks (Natick, USA, 2004). PLS Toolbox 6.2 from Eigenvector Research Inc. (Wenatchee, WA, USA) was used for building of PLS and classical LW-PLSR models and in-house written MATLAB functions were employed for modified LW-PLSR.

Table 1
Descriptive statistical parameters of the reference values of samples used throughout this study.

Analyte	Calibration set				Validation set			
	N	Mean conc.	SD	Interval	*N*	Mean conc.	SD	Interval
Proteins	332	6.5	0.7	4.1–9.3	331	6.5	0.7	4.1–8.6
Urea	584	48	36	15–249	583	48	36	15–242
Glucose	592	106	40	32–490	591	106	38	35–424
Triglycerides	510	127	84	51–1280	509	126	72	51–598

Note, *N*: number of samples and SD: standard deviation. All values are in mg/dL except for proteins which are in g/dL.