Chemometrics and Intelligent Laboratory Systems 154 (2016) 16–22

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Spectroscopic quantitation of tetrazolium formazan in nano-toxicity assay with interval-based partial least squares regression and genetic algorithm

Biswanath Mahanty, Soon-Uk Yoon, Chang-Gyun Kim *

Department of Environmental Engineering, INHA University, Incheon, 22212, Republic of Korea

ARTICLE INFO

Article history:
Received 4 January 2016
Received in revised form 9 March 2016
Accepted 10 March 2016
Available online 22 March 2016

Keywords:
Calibration
Formazan concentration
Genetic algorithm
Interval PLS

ABSTRACT

Spectrophotometric quantitation of formazan in tetrazolium-based nanoparticle (NP) toxicity assay requires a robust calibration model immune to optical interference. For the first time, variant of partial least squares (PLS) regression models, such as, full-spectrum (250–700 nm) PLS, interval PLS (*i*PLS), backward interval PLS (*bi*PLS), and synergy interval PLS (*si*PLS) models have been adopted for formazan quantitation. Models were evaluated based on root mean square error of cross-validation (RMSECV), and prediction (RMSEP). The spectral variables in optimal *i*PLS, *bi*PLS and *si*PLS models, as well as variables retained above a selection frequency threshold (for all intervals), were further refined in a genetic algorithm (GA). The results suggest that the optimal *bi*PLS (140 variables, 5 LVs, RMSECV: 0.4438, RMSEP: 0.2936) and *si*PLS (88 variables, 5 LVs, RMSECV: 0.4401, RMSEP: 0.316) models were superior either to the full-spectrum PLS (4 LVs, RMSECV: 0.9674, RMSEP: 0.4618) or traditional single wavelength calibration (414 nm, RMSECV: 2.0864, RMSEP: 2.1628). Minimum RMSEP (0.2976) was observed when GA was performed on spectral variables retained (above a threshold frequency) from the cumulative frequency distribution of all *si*PLS models. Finally, applicability of the selected PLS regression models in real NP toxicity assay is demonstrated.

1. Introduction

The extraordinary physicochemical properties of different nanoparticles (NPs) have driven their ever increasing applications ranging from catalysis, sensing, medical and environmental applications [1–3]. With the growing use of NPs, an understanding of their potential health hazards and ecological impact has become equally important [4,5]. Various cell viability-based *in vitro* cytotoxicity assays for NPs have been adopted in recent studies, where tetrazolium dye is enzymatically reduced to coloured formazan, presumably in proportion to the number of viable cells [6,7]. The produced formazan is solubilised in appropriate solvents and measured using a spectrophotometer at wavelengths usually between 450 and 600 nm.

Although the tetrazolium-based toxicity assay has been employed for a number of metal oxide NPs, the detection conditions described in the literature vary [8]. In addition, the intrinsic optical absorbance of NPs has been shown to interfere with spectroscopic measurements generating confounding or conflicting data [9–11].The influence of NPs on optical detection cannot be estimated in priori, which varies between samples depending on the composition [12]. Background-corrected single wavelength absorption of formazan can have little analytical value. In such case where corrections need to be made from the sample measurement itself, multivariate calibration methods such partial least squares (PLS) regressions provide excellent selectivity and prediction accuracy without requiring information on interfering analytes [13–16].

However, a full-spectrum PLS regression model may not be appropriate as a dimensionally reduced model with fewer selected variables would have better interpretation and prediction accuracy [17]. For local regression of spectral data, interval PLS (*i*PLS) is used to build PLS models on non-overlapping subintervals (often with different window sizes) on the spectrum to locate the optimum [18]. Synergy interval PLS (*si*PLS) with all possible combinations of intervals, or backward interval PLS (*bi*PLS) with sequential elimination of intervals offers significant improvements [19,20]. However, the subintervals retained in *bi*PLS or *si*PLS models may still contain much uninformative spectral variables. In the past decade, genetic algorithm-PLS (GA-PLS) has gained significant attention as a variable selection tool, which can be tuned to select variables from the full-spectrum regions as well as from spectral subintervals [21–24]. However, the numerous variables in the search domain reduce the capability of GA to find an optimal model, and number of variables can be reduced by applying the average of the signal intensities on a selected window size [22], or by using *i*PLS algorithm prior to GA implementation [19]. Each of the approaches has its own benefits and certain limitations, such that a generalization may not always be evident [25]. Variable selection methods find a 'good' set

* Corresponding author. Tel.: +82 32 860 7561.
E-mail address: cgk@inha.ac.kr (C.-G. Kim).

http://dx.doi.org/10.1016/j.chemolab.2016.03.012
0169-7439/

of variables rather than the 'optimal' and the applicability of these methods may be subjective [26].

In the present work, an interval-based PLS model is adopted for the spectrophotometric quantitation of formazan in iron oxide NP toxicity assay. The selected spectral regions from those interval-based models were processed in GA for further simplification, where their performance was assessed on the basis of the root mean square error (RMSE) of cross-validation (RMSECV), and prediction (RMSEP).

2. Materials and methods

2.1. Reagents and nanoparticle (NP) preparations

The iron oxide NPs were synthesised by the co-precipitation of ferrous and ferric ions in sodium hydroxide solution. The iron oxide NPs were subsequently coated with $Al(OH)_3$. The details for the preparation of the coated NPs and their characterization can be found elsewhere [27]. The functionalised iron oxide NP preparation (2 g/L in distilled water) was ultrasonically dispersed prior to its use. Other reagents, such as 2,3,5 triphenyl tetrazolium chloride (TTC), 1,3,5 triphenyl tetrazolium formazan (TPF), glucose, and 1,4-dioxane were obtained from Sigma Aldrich Chemical (Seoul, Korea). Concentrated hydrochloric acid (35–37%) was from Samchun Chemicals (Seoul, Korea).

2.2. Formazan calibration and nanoparticle (NP) toxicity assay

2.2.1. Formazan calibration sets

The formazan calibration sets were designed to mimic the composition of actual toxicity assay. *Escherichia coli* (ATCC 25922), the test microorganism in actual toxicity assay, was aerobically grown on sterile nutrient broth medium to an optical density of 0.73 (600 nm), centrifuged at 3500 × *g* for 15 min, and re-dispersed in distilled water. About 250 μL of the cell suspension was transferred into sets of 1.5 mL micro-centrifuge tubes. To each of the tubes, 50 μL of filter-sterilised glucose solution (1% w/v), 25–245 μL formazan solution (192 mg/L stock in 1,4 dioxane), 10–110 μL of TTC solution (336 mg/L stock), and varying amounts of NP suspensions (25, 50, 75, 100, or 125 μL from 2 g/L stock) were added (Table 1). Reaction volumes were adjusted to 435 μL with distilled water, the contents were mixed by vortexing, and finally 25 μL concentrated hydrochloric acid and 0.5 mL of 1,4-dioxane were added. This hydrochloric acid addition is not important for calibration samples, but in real toxicity assay it helps to release and dissolve intracellular formazan. The tubes were centrifuged at 3500 × *g* for 15 min, and the supernatants were collected for UV–Vis absorbance measurements.

2.2.2. Nanoparticle NP toxicity assay

NP toxicity assay was conducted in sets of 1.5 mL micro-centrifuge tubes. Cell suspension and glucose stock solution were added as described earlier for formazan calibration sets. Varying amounts of NP suspensions (25, 50, 75, 100, or 125 μL from 2 g/L stock) and 10 μL of TTC solution (5376 mg/L stock) were added to each of the tubes. The chosen TTC concentration ensured that the reduced formazan would be inside the concentration range in calibration sets, even when 100% TTC is microbially reduced. Reaction volumes were adjusted to 435 μL with distilled water, mixed by vortexing and incubated under dark at 25 °C for 45 min. Finally, formazan was extracted with 25 μL concentrated hydrochloric acid and 0.5 mL of 1,4-dioxane. The UV–Vis absorbance spectra of sample supernatants were recorded between 250 and 700 nm.

Table 1
Summary of the different calibration sets showing the number of samples (*N*), minimum (Min), maximum (Max), and mean of TTC or TTC-formazan and NP concentration.

Sets	*N*	Assay constituents and their concentration (mg/L)[a]								
		TTC-formazan			TTC			NPs		
		Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Calibration or validation	45	5	49	27	3.5	38.5	21	52.08	260.41	156.25

[a] Based on final assay mixture (960 μL:460 μL of aqueous phase and 500 μL of 1,4 dioxane).

2.3. Spectral data

The blank corrected UV–Vis absorbance of the supernatants for all samples was recorded between 250 and 700 nm at an interval of 1 nm (451 data points) using 1.0 cm path length cells. The independent variable matrix consisted of the assembly of spectral data of each sample as a row vector ($\mathbf{X}_{N\times 451}$) and the formazan concentration was utilised as the dependent variable ($\mathbf{Y}_{N\times 1}$). A test set ($N = 10$) and training set ($N = 35$) samples were selected by the Kennard–Stone algorithm [28]. The X matrix and Y vector were auto-scaled prior to PLS regression.

2.4. Model cross-validation and quality parameters

A PLS regression model extracts a number of uncorrelated latent variables (LVs) from measurements of inter-correlated variables to maximize the covariance between the **X** and **Y** data [29–31].To select the optimal number of LVs to be included, cross-validation was performed with five deletion groups created by taking every fifth sample from the concentration-sorted data matrix. Concentrations of samples from each deletion group were then validated against the model created with the remaining subsets of data. The process was repeated until each of the subsets was used exactly once as the validation data [32]. The number of LVs to be retained for a parsimonious model was determined with *F* statistical test, where the predicted error sum of squares (PRESS) value at the minimum is compared with PRESS values for all previous LVs. The number of LVs for the first PRESS values whose *F*-ratio probability drops below 0.90 was retained [31,33].

PLS model performances are quantitatively defined in terms of regression coefficient of CV (Q^2), RMSE of cross-validation (RMSECV), and RMSE of prediction (RMSEP) as follows:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (1)$$

$$RMSECV/P = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n}} \quad (2)$$

where n is the number of samples in the calibration (or prediction) set, y_i and $\tilde{y}_i$ are respectively the known and cross-validated (or prediction) estimate of ith sample, and $\bar{y}_i$ is the average of n observations.

2.5. Implementation of iPLS, biPLS, siPLS and GA-PLS strategy

2.5.1. Formulation of interval-based PLS models

For all interval-based models, the full-spectrum region is divided into 10, 11,…, 23, 25, 26, 28, 30, 32, and 34 intervals (total of 20 different cases, starting from about 45 variables/window to 14 variables/window). The framework for *i*PLS, *bi*PLS and *si*PLS model processing in GA is schematically shown in Fig. 1. For a given interval number, *i*PLS models are computed on each non-overlapping subintervals and the model with minimum RMSECV is retained. The *si*PLS and *bi*PLS models are based on forward selection and backward elimination of subintervals, respectively [20]. In *si*PLS, up to four subinterval combinations are considered while building a model. Starting from full-spectrum models, the *bi*PLS approach eliminates one subinterval at a time whose removal results in the lowest RMSECV [19]. In this study, backward elimination of subintervals is continued without any constrain (e.g., retaining a minimum number of variables). The whole process is