Bulletin of Faculty of Pharmacy, Cairo University xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Bulletin of Faculty of Pharmacy, Cairo University

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

Original Article

Improved predictions of nonlinear support vector regression and artificial neural network models via preprocessing of data with orthogonal projection to latent structures: A case study

Ibrahim A. Naguib*

Pharmaceutical Analytical Chemistry Department, Faculty of Pharmacy, Beni-Suef University, Alshaheed Shehata Ahmad Hegazy St., 62514 Beni-Suef, Egypt
Pharmaceutical Chemistry Department, Faculty of Pharmacy, University of Tabuk, Tabuk, Saudi Arabia

ARTICLE INFO

Keywords:
Agomelatine
SVR
ANN
OPLS
Spectrofluorimetry
Nonlinear

ABSTRACT

In the presented study, orthogonal projection to latent structures (OPLS) is introduced as a data preprocessing method that handles nonlinear data prior to modelling with two well established nonlinear multivariate models; namely support vector regression (SVR) and artificial neural networks (ANN). The proposed preprocessing proved to significantly improve prediction abilities through removal of uncorrelated data.

The study was established based on a case study nonlinear spectrofluorimetric data of agomelatine (AGM) and its hydrolysis degradation products (Deg I and Deg II), where a 3 factor 4 level experimental design was used to provide a training set of 16 mixtures with different proportions of studied components. An independent test set which consisted of 9 mixtures was established to confirm the prediction ability of the introduced models. Excitation wavelength was 227 nm, and working range for emission spectra was 320–440 nm.

The couplings of OPLS-SVR and OPLS-ANN provided better accuracy for prediction of independent nonlinear test set. The root mean square error of prediction RMSEP for the test set mixtures was used as a major comparison parameter, where RMSEP results for OPLS-SVR and OPLS-ANN are 2.19 and 1.50 respectively.

1. Introduction

In pharmaceutical analysis applications, nonlinearity of multivariate spectral data is an occasionally encountered problem that sometimes renders linear multivariate models unable to predict concentrations of drugs in their spectral data matrices in a proper way. This problem could arise due to constraints related to ratio of drug mixtures in drug products and related design levels that could force some of the levels to be at or sometimes beyond the border of linearity [1,2]. Accordingly, a group of nonlinear multivariate models were introduced to tackle the problem of nonlinearly and improve predictive abilities of analyzed multivariate test sets. Examples of these nonlinear models are nonlinear support vector regression (SVR) and nonlinear artificial neural networks (ANN) models; which proved through many previous applications to overcome the problem of nonlinearity [1].

Orthogonal projection to latent structures (OPLS) is considered a newly introduced data preprocessing method. It removes the systematic variations in the spectral data that is orthogonal to the concentration; resulting in easier interpretation of the results. The removed variations can be subjected to further analysis to get more knowledge about it [3].

The presented study utilizes the nonlinear spectrofluorimetric dataset created from analysis of agomelatine (AGM) and its hydrolysis degradation products (Deg I and Deg II), Fig. 1, [4], by a 3 factor 4 level experimental design resulting in a training set of 16 mixtures with different proportions of studied components. The concentration levels were selected carefully in a way that makes them outside the normal calibration range (at lower levels) and accordingly create a sort of nonlinearity in the relation between relevant concentrations and spectrofluorimetric response. An independent test set which consisted of 9 mixtures was established to test the prediction ability of the introduced models. The created training and test data sets were merely used a case study; to test the capabilities of chemometric models' combinations introduced.

The main aim of the presented study is to provide the chemometrics' society and pharmaceutical analysts with the results obtained from coupling OPLS preprocessing abilities to two well-established chemometric models (SVR and ANN) for analysis of nonlinear spectrofluorimetric data as a case study. The results of couplings affirm much improved prediction abilities of the nonlinear multivariate models; which favors the implementation of the coupling with OPLS for future

Peer review under responsibility of Faculty of Pharmacy, Cairo University.
* Address: Pharmaceutical Analytical Chemistry Department, Faculty of Pharmacy, Beni-Suef University, Alshaheed Shehata Ahmad Hegazy St., 62514 Beni-Suef, Egypt.
E-mail addresses: ibrahim.yaseeen@pharm.bsu.edu.sa, ibrahimnaguib1979@yahoo.com.

http://dx.doi.org/10.1016/j.bfopcu.2017.09.002
Received 28 March 2017; Received in revised form 27 July 2017; Accepted 9 September 2017

Please cite this article as: Naguib, I.A., Bulletin of Faculty of Pharmacy, Cairo University (2017), http://dx.doi.org/10.1016/j.bfopcu.2017.09.002

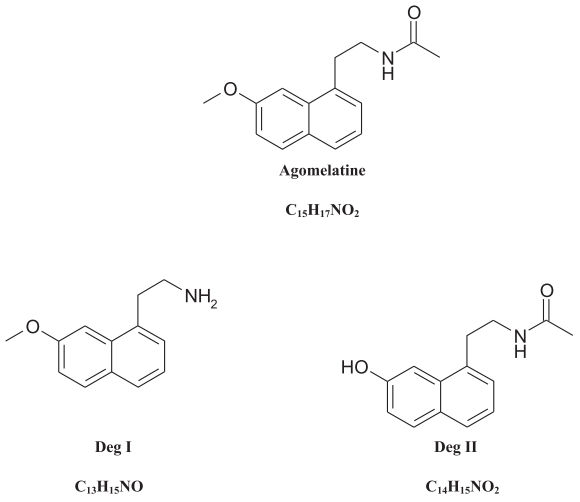

Fig. 1. Chemical structures of Agomelatine (AGM) and its hydrolysis degradation products (Deg I & II).

analysis of possibly encountered nonlinear data.

2. Experimental

2.1. Materials and reagents

- Agomelatine (AGM) pharmaceutical grade material was supplied by Inspire pharma company, Cairo, Egypt. The company's analysis certificate indicated 99.58% purity.
- Ethanol (HPLC grade; Fisher Scientific Ltd.).
- Deg I and Deg II were prepared according to literature [4].

2.2. Instruments

A Perkin-Elmer UK model LS-45 spectrofluorimeter, with a 150 W Xenon arc lamp and excitation and emission grating monochromators was used. A constant slit width of 10 nm was applied for both excitation and emission modes. The photomultiplier voltage was set to auto. 1% attenuation filter and a quartz 1 cm cuvette were used. The spectral band width was 0.5 nm and wavelength scanning speed was 1000 nm/min.

The spectra were collected using the software FL WINLAB, Version 4.00.03, Copyright 2006, Perkin Elmer, UK.

Excitation wavelength used was 227 nm, and spectra wehttps://ppts-suppliers.elsevier.com/pls/ppts_pl/pts1400w$.supplierre collected form 200–510 nm, with intervals of 0.5 nm. The working range for the presented study was 320–440 nm and the number of data points used in data analysis and modelling was 241 points, Fig. 2.

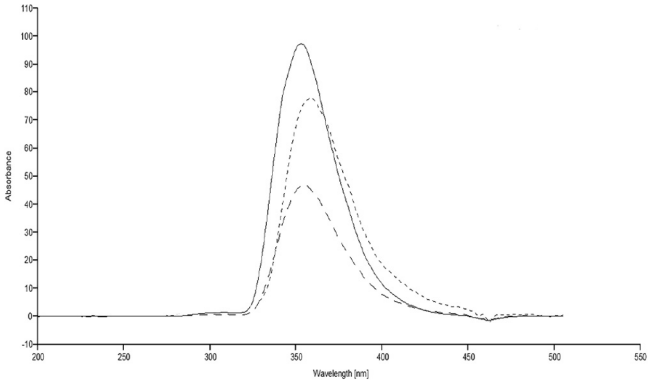

Fig. 2. Overlaid fluorescence emission spectra of 50 ng/ml AGM (—) and 15 ng/ml for each of Deg I (……) and Deg II (– – –) with excitation wavelength of 227 nm.

2.3. Software

Matlab® 7.1.0.246 (R14) was used for application of the different chemometric methods implemented in this study. The codes for the SVR algorithm were downloaded from the internet website (http://onlinesvr.altervista.org/, August 2009). Back-propagation neural network algorithm having three layers was implemented in Matlab using Neural Network toolbox. OPLS codes were written by H. Li and downloaded from MathWorks website (http://www.mathworks.com/matlabcentral/fileexchange/47767-libpls-1–95-zip/content/libPLS_1.95/opls.m, Jan 2016).

2.4. Experimental design for chemometric methods

The analyzed components; AGM and its degradation products (Deg I and II) were designed as a 3 factor 4 level calibration design using 4 concentration levels coded as –2, –1, +1 and +2 in which the level coded +1 was designated as the central level for each. The design aimed to cover the mixture space fairly well; ending up with 16 mixtures for the training set [5]. The central level of the design was 20 ng/ml, 1.8 ng/ml and 2.0 ng/ml for AGM, Deg I and II respectively. The concentration of the central level for AGM (20 ng/ml) was chosen to be at the lower border of its calibration range 20–100 ng/ml. The concentrations 16 and 14 ng/ml (levels -1 and -2 for AGM respectively) are practically out of calibration range.

Deg I and II concentration levels in the design were involved in about 7–11% calculated on molar basis from the central level of AGM. Mean centring proved to be the best data preprocessing method for obtaining optimum analysis results for used models. Nine independent test set mixtures were prepared to test the validity and predictive ability of the introduced chemometric models. These mixtures were prepared in such a way that they lay inside the mixture space of the training set, Fig. 3. Table 1 represents the concentration design matrix for both calibration and test sets. The working range for emission spectra was 320–440 nm and the number of data points used in data analysis and modelling was 241 points.

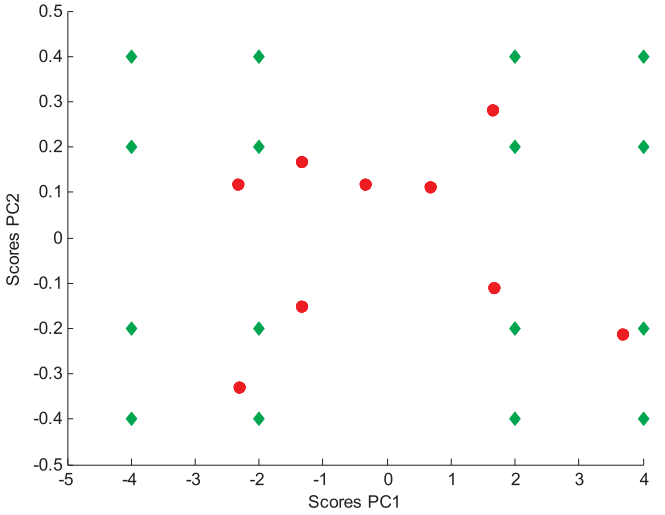

Fig. 3. Overlaid 2D Scores plot (PC1 Vs. PC2) for the mean centered concentration matrices of 16 training set mixtures (green diamonds) and 9 test mixtures (red circles) of the 4 level 3 factor experimental design. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)