Chemometrics and Intelligent Laboratory Systems 139 (2014) 168–174

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Application of genetic algorithms for pixel selection in multivariate image analysis for a QSAR study of trypanocidal activity for quinone compounds and design new quinone compounds

Maryam Sarkhosh [a], Neda Khorshidi [a], Ali Niazi [a,*], Riccardo Leardi [b]

[a] Department of Chemistry, Faculty of Science, Arak Branch, Islamic Azad University, Arak, Iran
[b] Department of Pharmaceutical and Food Chemistry and Technology, Genova University, Via Brigata Salerno (Ponte), Genova I-16147, Italy

ARTICLE INFO

Article history:
Received 30 May 2014
Received in revised form 24 July 2014
Accepted 8 September 2014
Available online 18 September 2014

Keywords:
QSAR
trypanocidal activity
multivariate image analysis
principal component regression
partial least squares
genetic algorithms

ABSTRACT

Quantitative structure-activity relationship (QSAR) analysis has been directed to a series of 31 quinone compounds with trypanocidal activity that was performed by chemometrics methods. The trypanocidal activity of the quinones is related to their redox potential (E_{pci}). Bidimensional images were used to calculate some pixels. Multivariate image analysis was applied to QSAR modeling of the redox potential of quinones derivatives by means of multivariate calibration such as principal component regression (PCR) and partial least squares (PLS). In this paper we investigate the effect of pixel selection by application of genetic algorithms (GAs) for PLS model. GAs is very useful in the variable selection in modeling and calibration because of the strong effect of the relationship between presence/absence of variables in a calibration model and the prediction ability of the model itself. The subset of pixels, which resulted in the low prediction error, was selected by genetic algorithm. The resulted model showed high prediction ability with RMSEP of 0.0694, 0.0358 and 0.0059 for PCR, PLS and GA-PLS models, respectively. Furthermore, the proposed QSAR model with GA-PLS was used for modification of structure and their activity predicted.

1. Introduction

Quantitative structure–activity relationship (QSAR) as one of the most important areas in chemometrics gives information useful for pharmaceutical chemistry, drug design, toxicology and eventually most facts of chemistry, and for this reason, several investigations have been carried out in order to improve the results [1–7]. The QSAR model is useful for understanding the factors controlling activity, for the prediction of activity and for designing new potent compounds. The main aim of QSAR studies is to establish an empirical rule or function relating the descriptors of compounds under the investigation of activities or properties. This rule of function is then utilized to predict the same activities of the compounds not involved in the training set from their descriptors. The activity that can be predicted with satisfactory accuracy depends to a great extent on the performance of the applied multivariate data analysis method, which has provided the property being predicted and is related to the descriptors. Model development in QSAR studies comprises different critical steps such as (1) descriptor generation, (2) data splitting to calibration (or training) and prediction (or validation) sets, (3) variable selection, (4) finding appropriate model between selected variables and activity and (5) model validation [7].

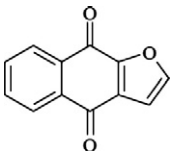
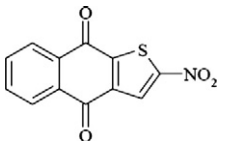
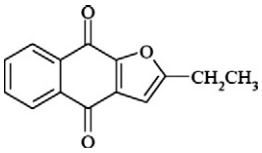
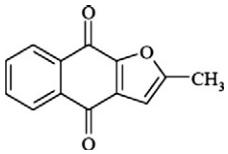
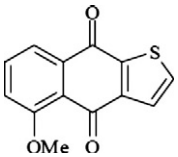
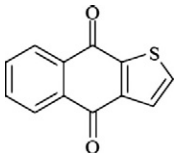
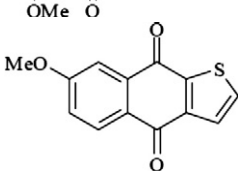
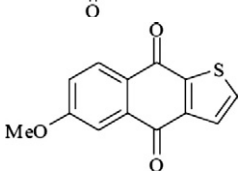
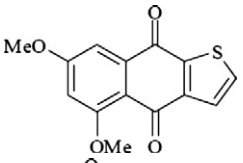
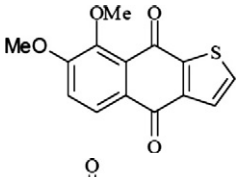
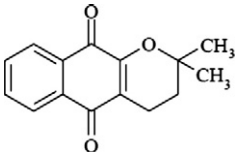
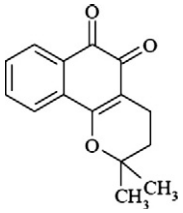
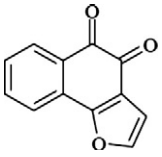
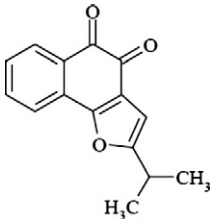
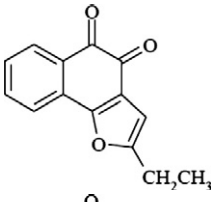
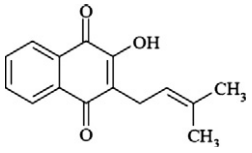
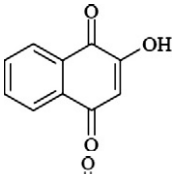
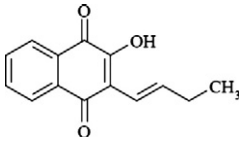
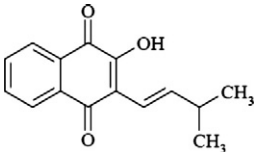
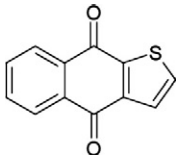
Among the investigation of QSAR, one of the most important factors affecting the quality of the model is a method to build the model. The traditional approach to QSAR relies heavily on multiple linear regression (MLR); however, due to the collinearity between descriptors, MLR is not able to extract useful information from data, and the overfitting problem is encountered [8]. Multivariate calibration such as PCR and PLS is a method that can be useful in dealing with the problem of the unfavorable more variable/object ratio and collinearity [9]. The PLS theory and its application in QSAR are reported by several of the workers [10–15]. Since it is not possible to know a priori which molecular properties are most relevant to the problem at hand, PLS, like other modeling methods, are often used in conjunction with optimization techniques for feature selection [16]. It has already been shown that genetic algorithms can be successfully used as a feature selection technique [17–26]. A GA is a stochastic method to solve optimization problems defined by a fitness criteria applying the evolution hypothesis of Darwin and different genetic functions, i.e., crossover and mutation. Leardi [27] demonstrated that GA, after suitable modifications, produces more interpretable results since the selected variables are less dispersed than with other methods.

A major step in constructing the QSAR models is finding molecular descriptors that represent variation in the structural property of the

* Corresponding author. Tel.: +98 912 5309767, fax: +98 86 33670017.
E-mail addresses: a-niazi@iau-arak.ac.ir, ali.niazi@gmail.com (A. Niazi).

http://dx.doi.org/10.1016/j.chemolab.2014.09.004
0169-7439/

Table 1
Chemical structure of quinones and their corresponding E_{pcl}.

Compound	E_{pcl}	Compound	E_{pcl}
	0.762*		0.413
	0.784		0.784
	0.832*		0.793
	0.847		0.817
	0.871*		0.855
	0.899		0.804
	0.659*		0.715
	0.681*		0.708*
	0.491		0.443
	0.405		0.359
	0.789		0.267

(continued on next page)