Chemometrics and Intelligent Laboratory Systems
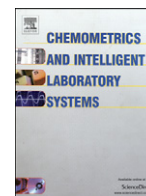
1 AUTHOR:

Maryam Sarkhosh

Islamic Azad University

14 PUBLICATIONS   **32** CITATIONS

CrossMark

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

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_{pcl}). 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.

© 2014 Elsevier B.V. All rights reserved.

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
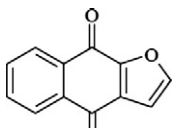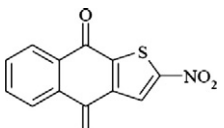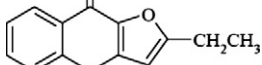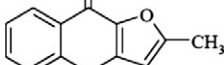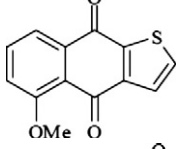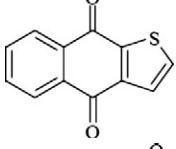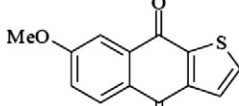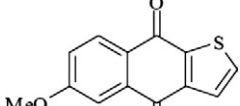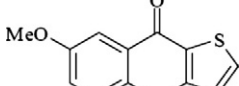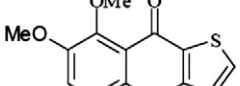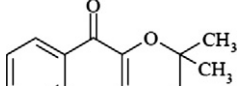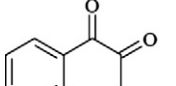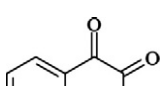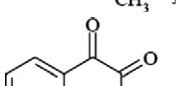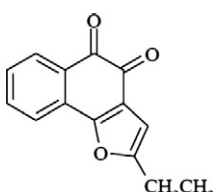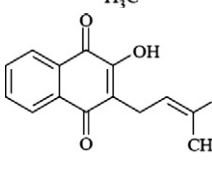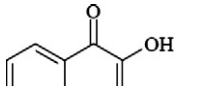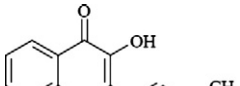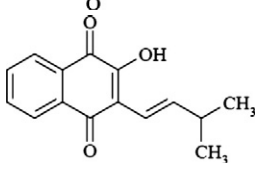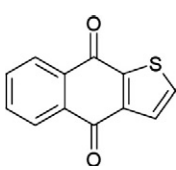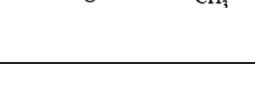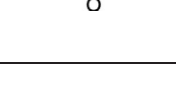
appropriate model between selected variables and activity and (5) model validation [7].

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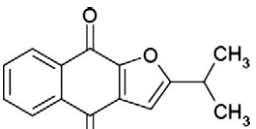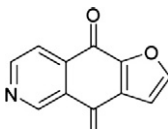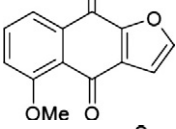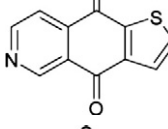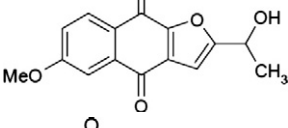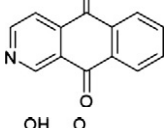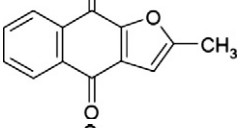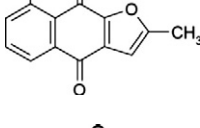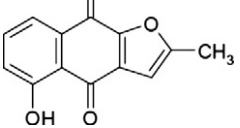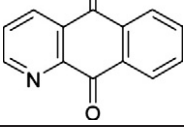
* Corresponding author. Tel.: +98 912 5309767, fax: +98 86 33670017.
  E-mail addresses: a-niazi@iau-arak.ac.ir, ali.niazi@gmail.com (A. Niazi).

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

Table 1 (continued)

Compound	E_{pcl}	Compound	E_{pcl}
(structure)	0.828[*]	(structure)	0.275
(structure)	0.817	(structure)	0.370[*]
(structure)	0.339	(structure)	0.365
(structure)	0.333	(structure)	0.359[*]
(structure)	0.370		

[*] Test set.

molecules by a number. Different descriptors have been studied to be used in QSAR analysis [28]. Nowadays, image analysis is becoming more important because of its ability to perform fast and non-invasive low-cost analysis on different processes in chemistry. Image analysis is a wide denomination that encloses classical studies on gray scale or (red–green–blue) RGB images [29]. Esbensen and Geladi [30] have demonstrated that image analysis may provide useful information in chemistry; the descriptors do not have a direct physicochemical meaning since they are binaries. In QSAR, images (2D chemical structure) have shown to contain chemical information [31–38], allowing the correlation between chemical structures and properties.

Quinones are compounds present in different families of plants; their molecular structures endow them with redox properties, which confer activity in various biological oxidative processes. The toxicity and therapeutic activities of these compounds involve the formation of oxygen reactive species. The biological redox cycle of quinones can be initiated by one electron reduction leading to the formation of semi-quinones, unstable intermediates that react rapidly with molecular oxygen generating free radicals [39]. Quinones are biologically important due to the case with which they are reduced to phenols. Such pairs of compounds serve as mediators of oxidation and reduction reactions in living organism.

The present study is focused on the application of 2D images, which are the proper structures of the compounds that can be drawn with aid of any appropriate program, as descriptors in QSAR. Then, multivariate image analysis-quantitative structure activity relationship study (MIA-QSAR) is proposed to model and predict the redox potential (E_{pcl}) as trypanocidal activity of a series of quinones by genetic algorithm-partial least squares (GA-PLS) modeling method.

As an application, the proposed method is tested for predicting the E_{pcl} of five new quinone compounds.

2. Materials and computational methods

2.1. Hardware and software

An HP personal computer (1 GB RAM) that was equipped with the Windows Vista operating system and MATLAB (Version 7.13, Mathwork Inc.) was used. The PLS evaluations were carried out by using the PLS program from PLS-Toolbox Version 2.0 for use with MATLAB from Eigenvector Research Inc. GA program was written in MATLAB by Leardi. The source codes of the programs are available from the authors upon our request and ChemOffice package (Version 2010) was used to draw the molecular structure. Kennard–Stones programs were written in MATLAB according to the algorithm [40].

2.2. Data set

The trypanocidal activities of different quinones were taken from literature [41–43]. The chemical structure of these quinones and their corresponding redox potential (E_{pcl}) values are shown in Table 1. In order to guarantee that training and prediction sets cover the total space occupied by the original data set, it was divided into the parts of training and prediction set according to the Kennard–Stones algorithm [40]. The Kennard–Stones algorithm is known as one of the best ways of building training and prediction sets and it has been used in many QSAR studies.
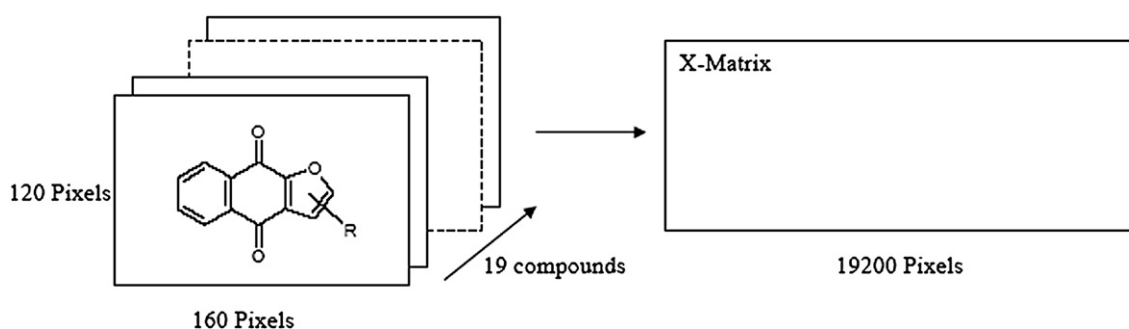
Fig. 1. 2D images and unfolding step of the 31 chemical structures to give the X-matrix. The arrow in structure indicates the coordinate of a pixel in common among the whole series of compounds, used in the 2D alignment step.
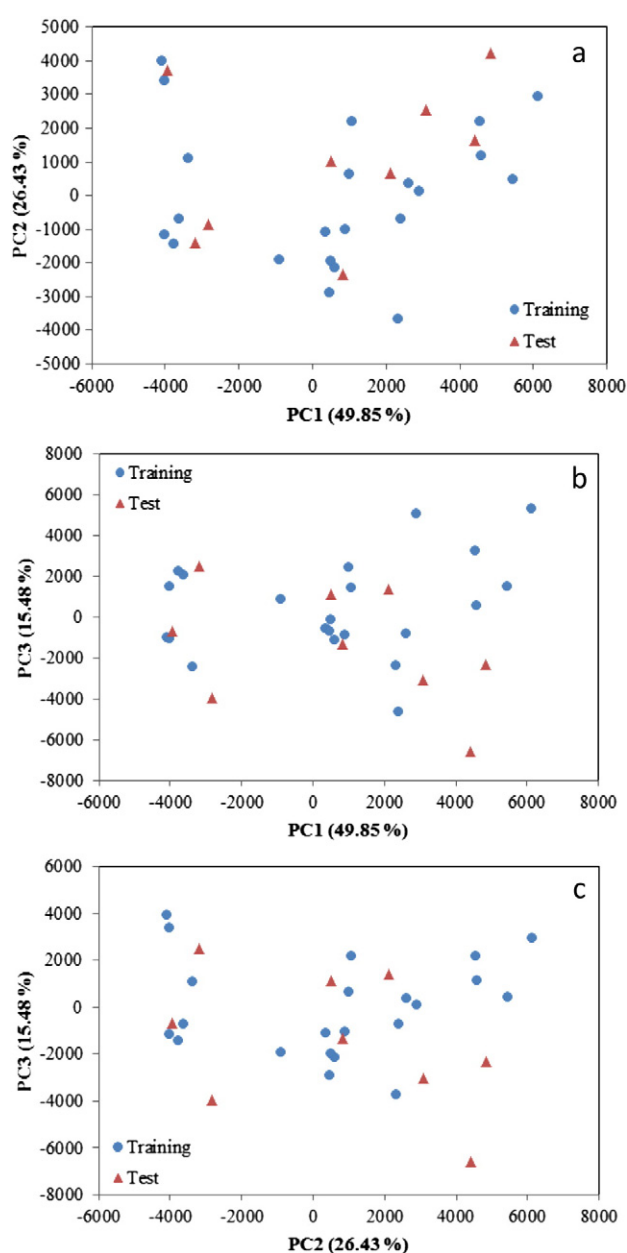
Fig. 2. Principal components analysis of the 2D image descriptors for the data set, (a) PC1 versus PC2, (b) PC1 versus PC3 and (c) PC2 versus PC3.
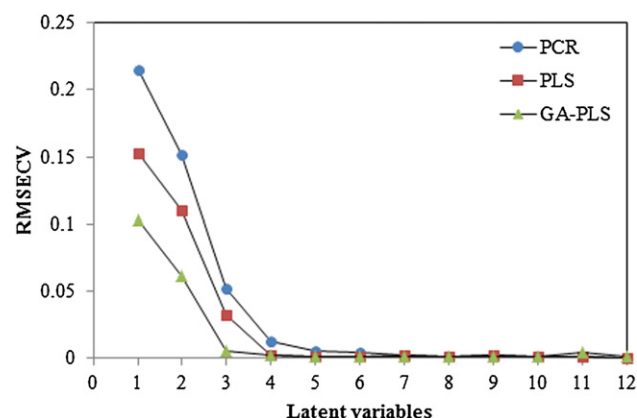
Fig. 3. The RMSECV versus number of latent variables.

2.3. Multivariate image analysis

In the MIA-QSPR method, the descriptors are the pixels of images that can be two or three dimensional. These pixels are correlated with dependent variables for making QSAR models. The 2D structures of each compound in Table 1 were systematically drawn in the ChemOffice program by same of font and size, and then, converted to bitmaps in 160×120 pixels workspace. Accordingly, pixels on the carbonyl carbons were fitted in the 80×20 pixel coordinate, making congruent the common structural scaffold of all structures; consequently, the variable structure responds for the variance in the Y block. Pixels were read as binaries in MATLAB, where black pixels are 0 and white pixels (where there is no chemical structure drawn) are 765, according to RGB color composition. Since each structure is a 2D image, the superposition (alignment) of the 31 images gives a three-way array of $31 \times 160 \times 120$ dimension, which was unfolded to a two-way array (matrix) of 19×19200 dimension. Many columns do not have variance because they correspond to either

Table 2
Parameters of the genetic algorithms.

Parameter [*]	Value
Population size	30 chromosomes
Response	cross-validated % explained variance
Maximum number of variables selected in the same chromosome	30
Probability of mutation	1%
Number of runs	100
Window size for smoothing	3

[*] On average, five variables per chromosome in the original population and backward elimination after every 100th evaluation and at the end.
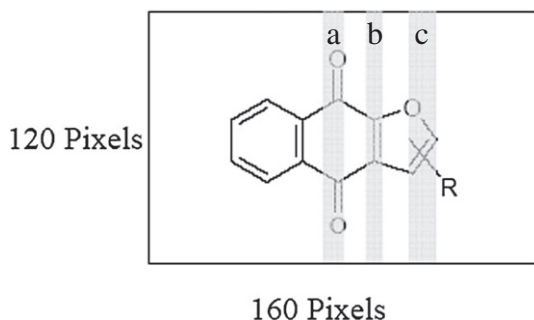
Fig. 4. Selected regions by genetic algorithms.

blank workspace or congruent structures, and therefore, they can be removed.

3. Results and discussion

3.1. Multivariate image analysis descriptors

The MIA-QSPR study is made according to the correlation of these pixels with dependent variables. The 2D structure of all quinones shown in Table 1, are drawn by ChemDraw program and then converted to bitmaps in 160´120 pixels workspace. All the drawn molecular structures were systematically fixed in a given coordinate. In this study, the pixel located at the 80´20 coordinate (carbonyl group), was used as a reference in the alignment step, as illustrated in Fig. 1. Each 2D image was read and converted into binaries (double array in MATLAB). Each image of dimension 160´120 pixels was unfolded at 19200 row and then the 31 images were grouped to form 31´19200 matrix. In order to minimize the memory, the columns with zero variance were reduced. As the result the matrix will is reduced in 31´5125 dimensional and all completely similar descriptors for all molecules are deleted and finally the number of descriptors is reduced to 174 then all pixels data are mean-centered.

3.2. Principal component analysis of the data set

In order to investigate collinearity and homogeneities in the data set and identify possible outliers and clusters, PCA were performed within the calculated image descriptors space for the whole data set. PCA is a useful multivariate statistical technique in which new variables (called principal components, PCs) are calculated as linear combinations of the old ones. These PCs are sorted by decreasing information content so that most of the information is preserved in the first few PCs. An important feature is that the obtained PCs are uncorrelated, and they can be used to derive scores that can be used to display most of the original variations in a smaller number of dimensions. These scores can also allow us to recognize groups of samples with similar behavior.

A total of 174 pixel were initially calculated by PCA for the entire data set of 31 compounds. The total number of pixels was reduced to 129 descriptors by eliminating the descriptors that were deemed insignificant (i.e. where the one-parameter correlation confinement with the activity is less than 0.1). The PCA results show that three PCs (PC1, PC2 and PC3) describe ~90% of the overall variances: PC1=49.85%, PC2= 26.43% and PC3=15.48% (Fig. 2). Since almost all variables can be accounted for the first three PCs, their score plot is a reliable presentation of the spatial distribution of the points for the data set. As can be seen from Fig. 2, there is not a clear clustering between compounds. The data separation is very important in the development of reliable and robust QSAR models. The quality of the prediction depends on the data set used to develop the mode. For regression analysis, data set was separated into two groups, a training set (22 data) and a prediction set (9 data) according to Kennard-Stones algorithm. As shown in Fig. 2, the distribution of the compounds in each subset seems to be relatively well-balanced over the space of the principal components.

Table 3
Observation and calculation values of E_{pcl} using PCR, PLS and GA-PLS models.

Number of compounds (Table 1)	Observation E_{pcl}	PCR		PLS		GA-PLS	
		Predicted	Error (%)	Predicted	Error (%)	Predicted	Error (%)
1	0.762	0.855	12.20	0.793	4.07	0.764	0.26
5	0.832	0.784	-5.77	0.824	-0.96	0.831	-0.12
9	0.871	0.752	-13.66	0.816	-6.31	0.862	-1.03
13	0.659	0.568	-13.81	0.611	-7.28	0.651	-1.21
15	0.681	0.742	8.96	0.698	2.50	0.685	0.59
16	0.708	0.723	2.12	0.762	7.63	0.709	0.14
23	0.828	0.893	7.85	0.845	2.05	0.831	0.36
26	0.370	0.391	5.68	0.389	5.14	0.375	1.35
30	0.359	0.322	-10.31	0.322	-10.31	0.353	-1.67
LVs		5		4		3	
RMSCEV		0.0051		0.0026		0.0018	
RMSEP		0.0694		0.0358		0.0059	
RSEP (%)		9.9440		5.1401		0.8572	

3.3. PCR and PLS modeling

The general purpose of the linear regression method is to quantify the relationship between several independent or predictive variables and a dependent variable. Independent or predictive variables could cause pixel changes in descriptors of image of molecules, their principal components or latent variables. The PCR and PLS methods are used to establish relationships between the dependent variables of the activity matrix and the pixels of the matrix as independent variables also that are called latent variables. The procedure performs a PCA on the independent variables matrix and simultaneously maximizing the correlation with the dependent variable matrix. In multivariate calibration such as PCR and PLS models, the cross-validation (leave-one-out) method was used for selecting the number of principal components and the predicted root mean square error cross-validation (RMSECV) is calculated [11,12]. The optimum number of factors was determined rather than the selection of the model, which yields a minimum in prediction error variance or RMSECV, the model selected is the one with the fewest number of factors such that RMSECV for that model is not significantly greater than the minimum RMSECV. One reasonable choice for the optimum number of factors would be that number which yielded the minimum RMSECV. Since there are a finite number of samples in the training set, in many cases the minimum RMSECV value causes overfitting for unknown samples that were not included in the model. A solution to this problem has been suggested by Haaland and Thomas [11,12] in which the RMSECV values for all previous factors are compared to the RMSECV values at the minimum. The F-statistical test can be used to determine the significance of RMSECV values greater than the minimum. As it is shown in Fig. 3, the RMSECV is minimized when the value of LVs is 5 and 4, thus, the optimum LVs for the training set of PCR and PLS methods was chosen to be 5 and 4, respectively. Prior to the PCR and PLS analysis, the data set was mean-centered.

3.4. GA-PLS modeling

As mentioned before, one of the problems is choosing the set of molecular (pixel) descriptors. GAs as an intelligent selection techniques [26], was utilized to fulfill this aim. Default values of the GAs program as written by Leardi were applied for most of the adjustable parameters of GAs, as listed in Table 2. All descriptors by mean-centering before performing the GA-PLS were performed. The GAs was optimized by variation and selection of the fitness values. After running of GAs for pixel selection, the selected pixels are used for running of PLS. When GA-

PLS is used the number of latent variables reduced to 3 (Fig. 3). The range selected pixel descriptors are shown in Fig. 4. The present study shows that the GAs can be a good method for pixel selection in image analysis. According to the selected descriptors by genetic algorithms it was found that the maximum structural effect are on E_{pcl} in a, b and c regions (Fig. 4) and among these three regions, c in which there are more changes by different functional groups influenced more than in other regions on E_{pcl}. As it is shown in Fig. 3, the RMSECV is minimized when the value of LVs is 3, thus, the optimum LVs for the training set of GA-PLS method was chosen to be 3.

3.5. Model validation and prediction of E_{pcl}

In Table 3, the predicted values of E_{pcl} obtained by the PCR, PLS and GA-PLS methods and the percent relative errors of prediction are presented. The correlation between observed and predicted E_{pcl} for GA-PLS was acceptable with $R^2 = 0.9990$ and intercept $= 0.0035$. The relative errors of prediction are between -1.67 to 1.35 % was obtained by using GA-PLS method. The data presented in Table 3 indicate that the GA-PLS model has good statistical quality with low prediction errors, while the GA-PLS model uses fewer latent variables. Also, for the evaluation of the predictive ability of a different model, the root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP) can be used:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^n \left(y_{i,pred} - y_{i,obs}\right)^2}{n}}$$

$$RSEP(\%) = 100 \times \sqrt{\frac{\sum_{i=1}^n \left(y_{i,pred} - y_{i,obs}\right)^2}{\sum \left(y_{i,obs}\right)^2}}$$

where $y_{i,pred}$ is the predicted E_{pcl} using different model, $y_{i,obs}$ is the observed value of the E_{pcl} and n is the number of compounds in the prediction set. Table 3 also shows RMSEP and RSEP for prediction of E_{pcl} of quinones. Other statistical parameters have been used for the evaluation of the suitability of the developed models for prediction of the activity of the studied compounds this include cross validation coefficient (Q^2 and R^2). These parameters are listed in Table 3 and show the good statistical qualities.

3.6. Molecular design

As an application of the proposed method, we investigated the GA-PLS model to predict the E_{pcl} of five new quinone compounds whose biological tests were not performed with them yet. Table 5 shows the chemical structure of five new quinone compounds and their E_{pcl} calculated by this proposed method. According to GA-PLS model, we have found that the new quinone molecule 5 (Table 4) is potentially active, i.e., it has a low redox potential. (See Table 5.)
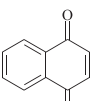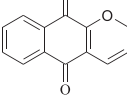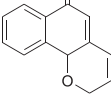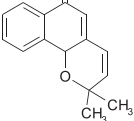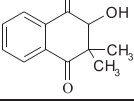
Table 4
Comparison of the statistical parameters by different QSAR models for the prediction of the E_{pcl}.

Methods	Data set	R^2	Q^{2*}
PCR	Training	0.8123	0.7955
	Test	0.7421	0.7128
PLS	Training	0.8563	0.8206
	Test	0.7548	0.7452
GA-PLS	Training	0.9256	0.9035
	Test	0.9154	0.9023

* Q^2 coefficient for the model validation by leave-one-out.

Table 5
Structural modification of quinone and predicted E_{pcl} by GA-PLS.

Number of design	Chemical structure	E_{pcl} calculated by GA-PLS
1		0.716
2		0.789
3		0.733
4		0.691
5		0.688

4. Conclusion

Using GA-PLS, a QSAR model for the prediction of E_{pcl} for 31 compounds based multivariate image analysis alone have been developed. The results well illustrate the power of pixels in prediction of E_{pcl} of quinones. The model could predict the E_{pcl} of quinone derivatives not existed in the modeling procedure accurately. The results of this study clearly show the potential and versatility of GA-PLS modeling in QSAR study of E_{pcl} of quinone derivatives using multivariate image analysis, which could be applied to prediction of E_{pcl}.

Conflict interest

There is no conflict interest.

Acknowledgments

References

[1] B. Hemmateenejad, R. Miri, M. Akhond, M. Shamsipur, Chemom. Intell. Lab. Syst. 64 (2002) 91–99.
[2] J. Mon, M. Flury, J.B. Harsh, J. Hydrol. 316 (2006) 84–97.
[3] A. Niazi, S. Jameh-Bozorghi, D. Nori-Shargh, Turk. J. Chem. 30 (2006) 619–628.
[4] A. Niazi, S. Jameh-Bozorghi, D. Nori-Shargh, J. Hazard. Mater. 151 (2008) 6063–6606.
[5] A. Coi, F.L. Fiamingo, O. Livi, V. Calderone, A. Martelli, I. Massarelli, A.M. Bianucci, Bioorg. Med. Chem. 17 (2009) 319–325.
[6] W. De-Eknamkul, K. Umehara, O. Monthakantirat, R. Toth, V. Frecer, L. Knapic, P. Braiuca, H. Noguchi, S. Miertus, J. Mol. Graph. Model. 29 (29) (2011) 784–794.
[7] M. Sarkhosh, J. Ghasemi, M. Ayati, Chem. Central J. 6 (2012) 1–8.
[8] L. Eriksson, J.L.M. Hermens, A multivariate approach to quantitative structure–activity relationships and structure–property relationships, in: J. Einax (Ed.), Chemometrics in Environmental Chemistry, the Handbook of Environmental Chemistry, vol. 5, Springer-Verlag, Berlin, Germany, 1995.
[9] P. Geladi, B.R. Kowalski, Anal. Chim. Acta. 185 (1986) 1–17.
[10] E.V. Thomas, D.M. Haaland, Anal. Chem. 62 (1990) 1091–1099.
[11] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193–1202.
[12] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1202–1208.
[13] Y. Liang, D. Yuan, Q. Xu, J. Chemom. 22 (2008) 23–35.

[14] G. Ioele, M.D. Luca, F. Oliverio, G. Ragno, Talanta 79 (2009) 1418–1424.
[15] T. Asadollahi, Sh. Dafarnia, A.M. Haji Shabani, J. Ghasemi, M. Sarkhosh, Molecules 16 (2011) 1928–1955.
[16] F. Ros, M. Pintore, J.R. Chretien, Chemometr. Intell. Lab. Syst. 63 (2002) 15–26.
[17] R. Leardi, J. Chemom. 15 (2001) 559–569.
[18] R. Leardi, R. Boggia, M. Terrile, J. Chemom. 6 (1992) 267–281.
[19] A. Broudiscou, R. Leardi, R. Phan-Tan-Luu, Chemom. Intell. Lab. Syst. 35 (1996) 105–116.
[20] R. Leardi, A. Lupianez Gonzalez, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.
[21] R. Leardi, J. Chemom. 14 (2000) 643–655.
[22] J. Ghasemi, A. Niazi, R. Leardi, Talanta 59 (2003) 311–317.
[23] A. Niazi, A. Soufi, M. Mobarakabadi, Anal. Lett. 39 (2006) 2359–2372.
[24] J. Ghasemi, D.M. Ebrahimi, L. Hejazi, R. Leardi, A. Niazi, J. Anal. Chem. 62 (2007) 348–354.
[25] J. Ghasemi, Sh. Ahmadi, Ann. Chim. 97 (2007) 69–83.
[26] A. Niazi, R. Leardi, J. Chemom. 26 (2012) 345–351.
[27] R. Leardi, Genetic algorithms in feature selection, in: J. Devillers (Ed.), In Genetic Algorithms in Molecular Modeling, Academic Press, London, 1996.
[28] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[29] J.M. Prats-Montalban, A. de Juan, A. Ferrer, Chemom. Intell. Lab. Syst. 107 (2011) 1–23.
[30] K. Esbensen, P. Geladi, J. Chemom. 3 (1989) 419–429.
[31] P. Geladi, Chemom. Intell. Lab. Syst. 9 (1990) 375–390.
[32] K. Esbensen, P. Geladi, H. Grahn, Chemom. Intell. Lab. Syst. 14 (1992) 357–374.
[33] M.P. Freitas, S.D. Brown, J. Martins, J. Mol. Struct. 738 (2005) 149–154.
[34] Z. Garkani-Nejad, M. Poshteh-Shirani, Talanta 83 (2010) 225–232.
[35] N. Khorshidi, M. Sarkhosh, A. Niazi, J. Sci. Innov. Res. 3 (2014) 189–202.
[36] M.P. Freitas, Chemom. Intell. Lab. Syst. 91 (2008) 173–176.
[37] C.A. Nunes, M.P. Freitas, J. Microbio. Methods 94 (2013) 217–220.
[38] C.A. Nunes, M.P. Freitas, Europ. J. Med. Chem. 62 (2013) 297–300.
[39] K.C.G. de Moura, F.S. Emery, C. Neves-Pinto, M.C. Pinto, A.P. Dantas, K. Salomão, S.L. de Castro, A.V. Pinto, J. Braz. Chem. Soc. 12 (2001) 325–338.
[40] R.W. Kennard, L.A. Stones, Technometrics 11 (1969) 137–148.
[41] J. Tonholo, L.R. Fretitas, F.C. Abreu, D.C. Azevedo, C.L. Zani, A.B. Oliveira, M.O.F. Goulart, J. Braz, Chem. Soc. 9 (1998) 163–169.
[42] J. Koyama, I. Morita, T. Yamori, Molecules 15 (2010) 6559–6569.
[43] M.O.F. Goulart, C.L. Zani, J. Tonholo, L.R. Freitas, F.C. de Abreu, A.B. Oliveira, D.S. Raslan, S. Starling, E. Chiari, Bioorg. Med. Chem. Lett. 7 (1997) 2043–2048.