Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Predicting biological activity of anticancer molecules 3-aryl-4-hydroxyquinolin-2-(1H)-one by DFT-QSAR models

Majdouline Larif
Faculty of Science, University Ibn Tofail - Kenitra
majdoulinelarif@yahoo.com

Samir Chtita
Faculty of Science, University Moulay Ismail, Meknes
samir.chtita@taalim.ma

Azeddine Adad
Faculty of Science, University Moulay Ismail, Meknes
azeddineadad@gmail.com

Rachid Hmamouchi
Faculty of Science, University Moulay Ismail, Meknes
r.hmamouchi@gmail.com

Mohammed Bouachrine
ESTM, University Moulay Ismail, Meknes, Morocco
bouachrine@gmail.com

Tahar Lakhlifi*
Faculty of Science, University Moulay Ismail, Meknes
tahar.lakhlifi@yahoo.fr

Abstract- Our objective is to study the relationship between the activities and structure, a 3D-QSAR study is applied to a set of 15 molecules for biological activity prediction derivatives. This study was conducted using the principal component analysis (PCA) method, the multiple linear regression method MLR and the artificial neural network (ANN). The predicted values of activities are in good agreement with the experimental results.

The PCA has allowed to observe the separation between two regions: Region 1 (λmax > 350nm; ΔE: 3.7 to 3.8) and region 2 (λmax < 330nm; ΔE: 4.1 to 4.3) and the distribution of different molecules of five groups.

Partial Least Square regression (PLS) techniques, considering the relevant descriptors obtained from the MLR, MNLR and ANN showed a correlation coefficient of 0.994 models which is a good result. As a result of quantitative structure–activity relationships, we found that the model proposed in this study is constituted of major descriptors used to describe these molecules. The obtained results suggested that the proposed combination of several calculated parameters could be useful to predict the biological activity of derivatives of 3-aryl-4-hydroxyquinolin-2-(1H)-one.

Keywords- Biological activity; 3D-QSAR model; RLM; ANN; PCA; DFT study; PLS

I. INTRODUCTION

This document is a template. An electronic copy can be downloaded from the Journal website. For questions on paper guidelines, please contact the journal publications committee as indicated on the journal website. Information about final paper submission is available from the conference website.

Abnormal fatty acid synthesis is among the most prevalent features of human cancer. Breast, prostate, colorectal, and ovarian cancers are among the list of common human tumors known to express high levels of fatty acid synthase (FAS, EC 2.3.1.85) and undergo constitutively elevated levels of fatty acid synthesis [1].

Since FAS is largely down-regulated by dietary fat in lipogenic tissues such as liver and adipose tissue, but remains highly expressed in many cancers, FAS is an attractive therapeutic target for cancer chemotherapy. Inhibition of FAS in vitro induces apoptosis in a variety of human cancer cells including breast, prostate, colon, and ovarian cell lines [2,3]. Treatment of human breast and prostate cancer xenografts in athymic mice with FAS inhibitors have shown a significant anti-tumor effect without toxicity to proliferating normal tissues such as bone marrow, skin, liver, and gastrointestinal tract [1,4].

Quantitative structure-activity relationship (QSAR), as an important area of chemometrics, has been the subject of a series of investigations [5]. The main aim of QSAR studies is to establish an empirical rule or function relating the structural descriptors of compounds under investigation to bioactivities. This rule or function is then utilized to predict the same bioactivities of the compounds not involved in the training set from their structural descriptors. Whether the bioactivities can be predicted with satisfactory accuracy depends to a great extent on the performance of the applied multivariate data analysis method, provided the property being predicted is related to the descriptors. Many multivariate data analysis methods such as principal components analysis (PCA), partial least square regression (PLS) and artificial neural network (ANN) have been used in QSAR studies. PLS, as a most commonly used on chemometrics method, has been extensively applied to QSAR investigations. ANN and PLS offers satisfactory accuracy in most cases but tends to over fit the training data. There are a large number of molecular descriptors that can be used in QSAR studies. Once validated, the findings can be used to predict activities of untested compounds. Recently, computer-assisted drug design based on QSAR has been successfully employed to develop new drugs for the treatment of cancer and other diseases [6].

After a QSAR model is built and validated, it can predict the biological activity of novel molecules room their structural properties. A QSAR model can also screen potentially active molecules from a database, as described in the section on applications of the technique. Because the QSAR model can incorporate a wide range of different variables, be it physical, chemical or biological, it can also be utilized in industries apart from drug design [7], such as toxicology [8], food chemistry [9], and other fields.

Considering the increasing interest in new anticancer compounds and the potential of the FAS as a target for anticancer drug design, the main goal of this work is to provide a SAR and a QSAR (quantitative structure-activity relationships) studies of a selected training set of 3-aryl-4-hydroxyquinolin-2-(1H)-one derivatives (compounds structurally related to flavonoids, (Fig. 1) available in the literature [10] and that presented ability to inhibit FAS.

II. MATERIAL AND METHODS

A. Material

All compounds studied in this work are 3-aryl-4-hydroxyquinolin-2-(1H)-one derivatives. This data set, containing fifteen compounds (Table 1), one of the biggest sets of human FAS type I inhibitors available in literature, was described and the biological activity was investigated by Rivkin and co-workers [10]. Therefore, it was selected for this study.

The biological activity was measured according to the concentration (in nanomolar, nM) required for fifty percent inhibition of the Human FAS enzymatic activity, pI_{50}. The observed pI_{50} measurements were converted into their corresponding log IC_{50} (or pIC_{50}) and are also presented in table 1. The following figure **1a** represents the basic strucutre flavonoid 3-aryl-4-hydroxyquinoline-2-(1H)-one and figure **1b** below shows the chemical structures of the compounds studied and experimental activities corresponding FAS and pI_{50}.

Fig.1a: Basic structure of flavonoids and 3-aryl-4-hydroxyquinolin-2-(1H)-one

Fig.1b: Structures of flavonoids and 3-aryl-4-hydroxyquinolin-2-(1H)-one.

TABLE 1: SELECTED TRAINING SET OF 3-ARYL-4-HYDROXYQUINOLIN-2-(1H)-ONE DERIVATIVES AND THEIR RESPECTIVE POTENCIES REFERENT TO HUMAN FAS TYPE INHIBITORS [10]

Molecule	FAS	pI_{50}
1	1,403	5,853
2	207	6,684
3	1,294	5,888
4	136	6,866
5	402	6,396
6	52	7,284
7	4,164	5,380
8	54	7,268
9	68	7,167
10	728	6,138
11	127	6,896
12	101	6,996
13	19	7,721
14	734	6,134
15	194	6,712

B. Computational methods

- DFT calculations

DFT (density functional theory) methods were used in this study. These methods have become very popular in recent years because they can reach similar precision to other methods in less time and less cost from the computational point of view. In agreement with the DFT results, energy of the fundamental state of a polyelectronic system can be expressed through the total electronic density, and in fact, the use of electronic density instead of wave function for calculating the energy constitutes the fundamental base of DFT [11-13], using the B3LYP functional [14,15] and a 6-31G* basis set. The B3LYP, a version of DFT method, uses Becke's three-parameter functional (B3) and includes a mixture of HF with DFT exchange terms associated with the gradient corrected correlation functional of Lee, Yang and Parr (LYP). The geometry of all species under investigation was determined by optimizing all geometrical variables without any symmetry constraints.

- Calculation of molecular descriptors using Gaussian 03W

From the results of the DFT calculations, the quantum chemical descriptors were obtained for the model building as follows: the total energy (E_T (u.a.)), the highest occupied molecular orbital energy (E_{HOMO} (eV)), the lowest unoccupied molecular orbital energy (E_{LUMO} (eV)), the energy difference between the LUMO and the HOMO energy (E_{GAP} (eV)= ΔE), absorption maximum λmax, the total dipole moment of the molecule (μ (Debye)), absolute hardness (η), absolute electron negativity (χ) and reactivity index (ω) [16]. η, χ and ω were determined by the following equations:

$$\eta = \frac{(E_{LUMO} - E_{HOMO})}{2} \quad ; \quad \chi = -\frac{(E_{LUMO} + E_{HOMO})}{2} \quad ; \quad \omega = \frac{\chi^2}{2\eta}$$

- Principal components analysis

Fifteen molecules were studied by statistical methods based on the principal component analysis (PCA) [17,18] using the software XLSTAT 2009 and Matlab software v 2009a.

This is an essentially a descriptive statistical method which aims to present, in graphic form, the maximum of information contained in the data table 1.

PCA is a statistical technique useful for summarizing all the information encoded in the structures of compounds. It is also very helpful for understanding the distribution of the compounds.

- Multiple linear regressions (MLR)

The multiple linear regression statistic technique is used to study the relation between one dependent variable and several independent variables. It is a mathematic technique that minimizes differences between actual and predicted values. The multiple linear regression model (MLR) was generated using the software SYSTAT, version 12, to predict antiamoebic activities $logIC_{50}$. It has served also to select the descriptors used as the input parameters for a back propagation network (ANN).

- Partial least square analysis (PLS)

The PLS have two objectives: to approximate the matrix X of molecular structure descriptors to the matrix Y of dependent variables and to maximize the correlation between them. The leave-one-out (LOO) method [19] was used to perform the cross-validated analysis.

The optimal number of components (N) is employed to do non-validation PLS analysis to get the final model parameters such as correlation coefficient R^2 [20], standard deviation (S) and Fischer test value (F).

- Artificial neural networks (ANNs)

The ANNs analysis was performed with the use of Matlab software v 2009a Neural toolbox on a data set of structures of flavonoids and 3-aryl-4-hydroxyquinolin-2(1H)-one. [21,22]. A number of individual models of ANN were designed built up and trained. Generally the network was built for three layers; one input layer, one hidden layer and one output layer were considered [23]. The input layer was consisted of eight artificial neurons of linear activation function (Fig.2). The number of artificial neural in the hidden layer was adjusted experimentally. The hidden layer consisted of 15 artificial neural. One neuron formed the output layer of sigmoid function activation. The architecture of the applied ANN models is presented in figure 3.

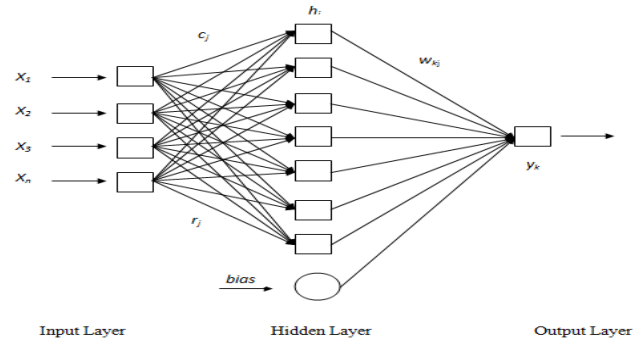

Fig. 2: Neuron Layout of ANN

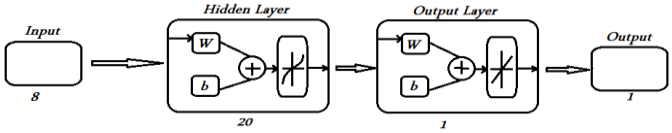

Fig. 3: The ANN architecture.

The data subjected to ANN analysis were randomly divided into three sets: a learning set, a validation set and a testing set. Prior to that, the whole data set was scaled within the 0-1 range.

The set of structures of flavonoids and 3-aryl-4-hydroxyquinolin-2-(1H)-one [24] was subjected to the ANN analysis. First, for the learning set of compounds, i.e., Selected training set of 3-aryl-4-hydroxyquinolin-2-(1H)-one derivatives and their respective potencies referent to human FAS type I inhibitors. The learning set of data is used in ANN to recognize the relationship between the input and output data. Then for the revision of the ANN model designed and selected, the validation set of 15 compounds was used. Testing set with eight compounds was provided to be an independent evaluation of the ANN model performance for the finally applied network.

In this study, we selected the sigmoid as a basis function [25].

The operation of the output layer is linear, which is given as below:

$$y_k(X) = \sum_{j=1}^{n_k} w_{kj} h_j(X) + b_k$$

Where y_k is the k_{th} output layer unit for the input vector X, w_{kj} is the weight connection between the k_{th} output unit and the j_{th} hidden layer unit and b_k is the bias that allows a transfer function "non-zero" given by the following equation:

$$\text{Bias} = \sum(\bar{y} - y)$$

Where y is the measured value and $\bar{y}$ is the value predicted by the model.

The accuracy of the model was mainly evaluated by the root mean square error (RMSE). Formula is given as follows:

$$RMSE = \sqrt{\frac{1}{n}.\sum_{i=1}^{n}(p_{exp} - p_{pred})^2}$$

Where n = number of compounds, p_{exp} = experimental value, p_{pred} = predicted value and summation is of overall patterns in the analyzed data set [26,27]. The scripts were run on a personal PC.

III. RESULTS AND DISCUSSION

A QSAR study was carried for a series of 15 of 3-aryl-4-hydroxyquinolin-2-(1H)-one derivatives and their respective potencies referent to human FAS type I inhibitors, in order to determine a quantitative relationship between structure and toxicity

Table 2 shows the values of the calculated parameters obtained by DFT/B3LYP 6-31G* optimization of the studied Selected training set of 3-aryl-4-hydroxyquinolin-2-(1H)-one derivatives and their respective potencies referent to human FAS type I inhibitors.

TABLE 2: VALUES OF THE THIRTEEN CHEMICAL DESCRIPTORS

N°	FAS	pI_{50}	Et (u.a.)	E_{HOMO} (eV)	E_{LUMO} (eV)	ΔE (eV)	μ (D)	χ	η	ω	Ea(eV)	λ_{max} (nm)	$f_{(SO)}$
1	1,40	5,853	-26901,945	-6,301	-2,403	3,899	3,431	4,352	1,949	4,858	3,446	359,76	0,037
2	207,00	6,684	-39416,309	-6,391	-2,515	3,876	4,298	4,453	1,938	5,116	3,409	363,68	0,0281
3	1,29	5,888	-23845,139	-6,219	-1,880	4,339	3,145	4,049	2,170	3,779	3,874	320,05	0,2668
4	136,00	6,866	-36359,789	-6,325	-2,059	4,266	4,280	4,192	2,133	4,120	3,823	324,3	0,322
5	402,00	6,396	-36359,842	-6,380	-2,089	4,291	3,458	4,235	2,145	4,179	3,823	324,34	0,2512
6	52,00	7,284	-42651,481	-6,248	-2,046	4,202	4,358	4,147	2,101	4,092	3,759	329,83	0,2747
7	4,16	5,380	-58659,604	-6,467	-2,213	4,254	3,648	4,340	2,127	4,428	3,803	326,02	0,3885
8	54,00	7,268	-45353,597	-6,328	-2,093	4,235	4,788	4,211	2,118	4,186	3,789	327,2	0,3285
9	68,00	7,167	-48055,516	-6,363	-2,134	4,229	4,893	4,248	2,114	4,268	3,784	327,68	0,3013
10	728,00	6,138	-48055,695	-6,311	-2,202	4,110	2,929	4,256	2,055	4,408	3,820	324,58	0,3722
11	127,00	6,896	-46835,197	-5,881	-2,055	3,826	4,729	3,968	1,913	4,115	3,462	358,18	0,0034
12	101,00	6,996	-55569,557	-5,880	-2,092	3,787	4,794	3,986	1,894	4,195	3,397	364,97	0,0071
13	19,00	7,721	-52061,532	-5,731	-2,002	3,728	5,566	3,867	1,864	4,010	3,394	365,34	0,01
14	734,00	6,134	-49848,999	-5,875	-2,063	3,812	2,906	3,969	1,906	4,133	3,459	358,4	0,0005
15	194,00	6,712	-43088,129	-6,416	-2,145	4,270	2,225	4,281	2,135	4,291	3,818	324,74	0,3654

A. Principal component analysis (training set selection)

The selection of the training set is one of the most important steps in the QSAR modeling, since the establishment and optimization of a QSAR model are based on this training set. Predictability and applicability of a QSAR model also depend on the training set selection. In this part, PCA was applied to select a training set from among 15 compounds.

The set of descriptors encoding the 15 compounds and electronic and energetic parameters are submitted to PCA analysis (XSLAT 2009). The first two principal axes are sufficient to describe the information provided by the data matrix. Indeed, the percentages of variance are 44.87% and 25.37% for the axes F1 and F2 respectively. The total information is estimated to a percentage of 70.24%. The principal component analysis (PCA) [22,28,29] was conducted to identify the link between the different variables. Bold values are different from 0 at a significance level of p = 0.05.

Table 3 shows the descriptor's contributions to F1 and F2.

TABLE 3: DESCRIPTOR'S CONTRIBUTIONS TO THE FIRST TWO PRINCIPAL COMPONENTS F1 AND F2.

Descriptors	F1		F2	
	Correlations	Contributions %	Correlations	Contributions %
FAS	0,052	0,047	0,227	1,559
pI_{50}	-0,546	5,119	-0,257	1,995
Et	0,463	3,670	-0,333	3,360
E_{HOMO}	-0,795	10,829	-0,388	4,570
E_{LUMO}	-0,386	2,555	-0,913	25,261
ΔE	0,893	13,674	-0,391	4,634
μ	-0,605	6,274	-0,043	0,056
χ	0,688	8,108	0,668	13,520
η	0,893	13,674	-0,391	4,634
ω	0,360	2,216	0,924	25,877
Ea	0,833	11,902	-0,488	7,233
λmax	-0,833	11,902	0,488	7,233
$f_{(SO)}$	0,765	10,033	0,047	0,067

Correlations between the thirteen descriptors are shown in table 3 as a correlation matrix, in figures 4 and 5 these descriptors are represented in a correlation circle.

The Pearson correlation coefficients are summarized in the following table 4. The obtained matrix provides information on the negative or positive correlation between variables.

TABLE 4: CORRELATION MATRIX (PEARSON (N)) BETWEEN DIFFERENT OBTAINED DESCRIPTORS

Variables	FAS	pI_{50}	E_T	E_{HOMO}	E_{LUMO}	ΔE	μ	χ	η	ω	E_a	λ_{max}	$f_{(SO)}$
FAS	**1**												
pI_{50}	-0,036	**1**											
E_T	-0,129	-0,193	**1**										
E_{HOMO}	-0,118	0,279	-0,175	**1**									
E_{LUMO}	-0,196	0,411	0,096	**0,657**	**1**								
ΔE	-0,118	-0,311	**0,543**	**-0,632**	0,007	**1**							
μ	-0,332	**0,761**	-0,314	0,232	0,225	-0,418	**1**						
χ	0,043	-0,471	0,257	**-0,832**	**-0,882**	0,371	-0,350	**1**					
η	-0,118	-0,311	**0,543**	**-0,632**	0,007	**1,000**	-0,418	0,371	**1**				
ω	0,204	-0,418	-0,118	**-0,632**	**-0,996**	-0,036	-0,221	**0,868**	-0,036	**1**			
E_a	0,075	-0,343	0,429	-0,446	0,121	**0,904**	-0,496	0,196	**0,904**	-0,146	**1**		
λ_{max}	-0,075	0,343	-0,429	0,446	-0,121	**-0,904**	0,496	-0,196	**-0,904**	0,146	**-1,000**	**1**	
$f_{(SO)}$	-0,125	-0,143	0,036	**-0,700**	-0,350	**0,664**	-0,236	**0,561**	**0,664**	0,318	**0,654**	**-0,654**	**1**

Bold values are different from 0 at a level significant for $p < 0.05$
At a very significant for $p < 0,01$
At a highly significant to $p < 0,001$

Correlation circle

Principal component analysis (PCA) was also performed to detect the connection between the different variables. The principal component analysis revealed from the correlation circle (Fig.4) shows that the F1 axis (44.87%) presents the energy E_{LUMO} of the variance while the axis F2 (25.37%) of the variance is located by the other parameters of energy.

The toxicity pI_{50} is well correlated positively with the for r = 0,761 and $p < 0.05$ at a significant level.

- ΔE and η are perfectly correlated (r = 1), both variables are redundant.
- E_{LUMO} and ω are strongly negatively correlated (r = -0.996).
- Ea and λmax are strongly negatively correlated (r = -1).

The following variables then removed are: **η,□ λ_{max},** and **ω□**

On the other hand, the correlation circle (Figure) indicates the correlation between electronic descriptors:

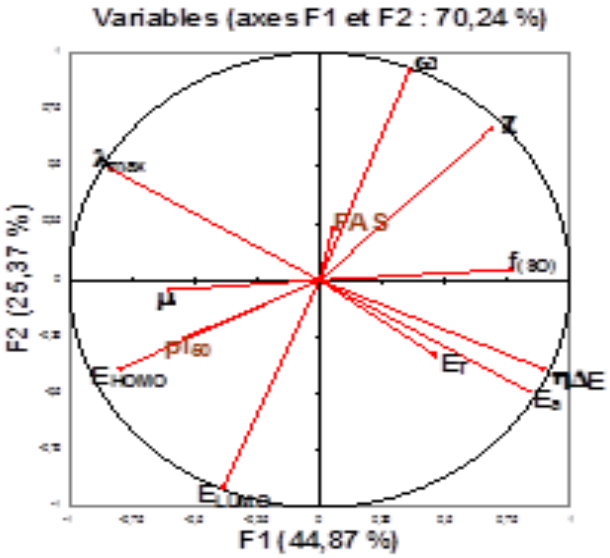

Fig 4: Correlation circle

The Cartesian diagram (Figure 5) Analysis of projections according to the plane F1–F2 (70.24%) of the total variance of the studied molecules. We notice that the pI_{50} toxicity is strongly correlated with E_{HOMO}

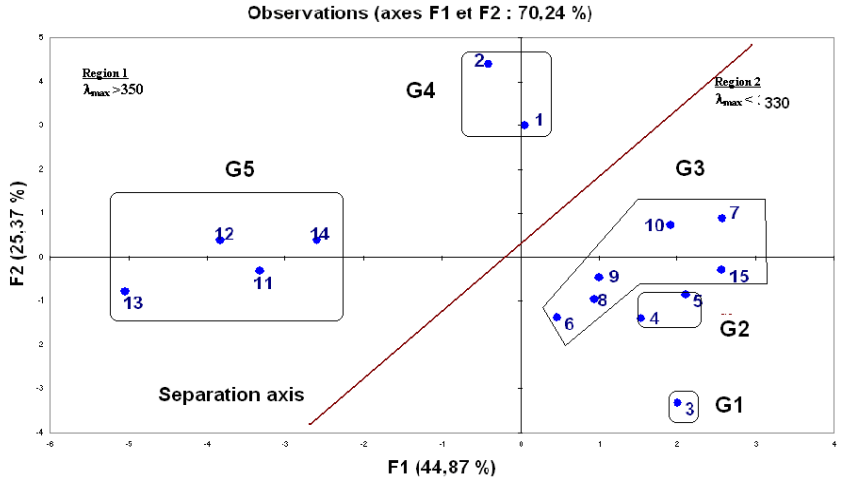

Fig. 5: Cartesian diagram showing the separation between the two regions and the dispersal of different molecules by groups.

On the Cartesian diagram we can distinguish two regions:

Region 1 (λ_{max} > 350; ΔE : 3.7 to 3.8) and region 2 (λ_{max} < 330 ; ΔE : 4.1 to 4.3) separated by a separation axis and different groups of molecules (table 2):

G1 (molecule 3 which includes a C≡N group in R^1 and R^2 no Cl (Figure 1a);

G2 (Contains two positional isomers and R^5 = H (figure 1a) ;

G3 (Molecules containing variously substituted aromatic R^5 (Figure 1a) ;

G4 (containing molecules $R^1 = NO_2$);

G5 (molecules containing aryl benzocondensed R^5 and biphenyl-type (Figure 1a).

B. Multiple Linear Regressions (MLR)

In order to propose a mathematical model and to evaluate quantitatively the substituent's physicochemical effects on the activity pI_{50} of the totality of the set of these 15 molecules, we submitted the data matrix constituted obviously from the 13 physicochemical variables corresponding to the 15 molecules, to a progressive multiple regression analysis. This method used the coefficients R, R^2, and the F-values to select the best regression performance.

Where R is the correlation coefficient; R^2 is the coefficient of determination; MSE is the mean squared error; F is the Fisher F-statistic.

Treatment with multiple linear regressions is more accurate because it allows you to connect the structural descriptors for each activity of 15 molecules to quantitatively evaluate the effect of substituent. The selected descriptors are:

To linearly correlate the molecule descriptors: the total energy **Et**, activation energy **Ea**, energy $\mathbf{E_{GAP}}$ **(eV)= ΔE**, energy $\mathbf{E_{HOMO}}$, energy $\mathbf{E_{LUMO}}$, the dipole moment **μ**, absorption maximum **χ** and the factor of oscillation $\mathbf{f_{(SO)}}$ to $\mathbf{pI_{50}}$, the following equation was used:

$$\mathbf{pI_{50}} = 9{,}6263 + 0{,}8205.\mathbf{E_{HOMO}} + 0{,}4728.\boldsymbol{\mu} + 1{,}1115.\mathbf{f_{(SO)}} \quad \text{(equation 1)}$$

N = 15 R = 0,744 R^2 = 0,554 RMSE = 0,487 MCE= 0,237

The equation 1 depends only on tree descriptors, the coefficient of correlation is 0.744 is an acceptable result. We do not get a good alignment of the molecules along the line of (Figure 6).

TABLE 5: ANALYSIS OF VARIANCE

Source	DDL	Sum of squares	Mean square	F	Pr > F
Model	3	3,236	1,079	4,550	0,026
Error	11	2,607	0,237		
Total corrected	14	5,843			

As a remark, the model the values are different from 0 at a significant level p <0.05 for Pr <0,026 with $F_{(3,14)}$ = 4,550

The figure 6 shows a very regular distribution of toxicity values depending on the experimental values.

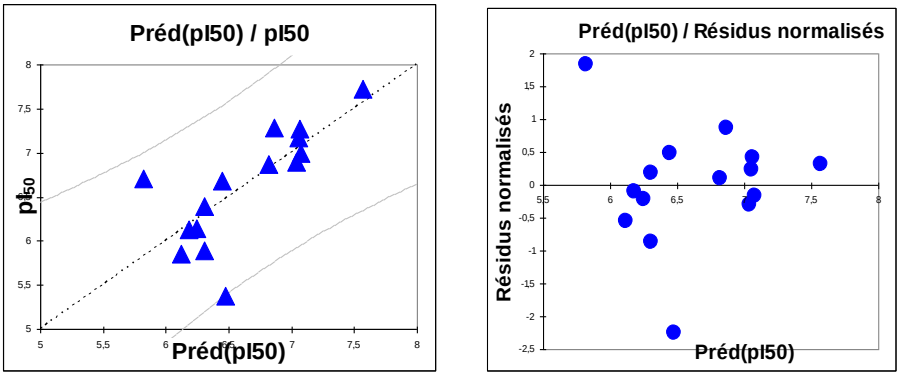

Fig. 6: Relationship between the estimated values of pI_{50}, their predictions and their residues established by MLR

C. Multiples non linear regression (MNLR)

We have used also the technique of nonlinear regression model to improve the structure - activity relationship to quantitatively evaluate the effect of substituent. It takes into account several parameters. This is the most common tool for the study of multidimensional data. We have applied to the data matrix constituted obviously from the descriptors proposed by MLR corresponding to the 15 molecules. The coefficients R, R^2, and the F-values are used to select the best regression performance.

$$\mathbf{pI_{50}} = -129{,}13048 - 1{,}9114.10^{-4}.\mathbf{Et} - 641{,}3908.\mathbf{E_{HOMO}} + 230{,}2484.\mathbf{E_{LUMO}} - 404{,}8066.\mathbf{\Delta E} - 1{,}19212.\mu\square - 370{,}5049.\chi\square - 2{,}7454.\mathbf{E_a} + 10{,}6399.\mathbf{f_{(SO)}} - 2{,}39506\ 10^{-9}.\mathbf{Et}^2 - 27{,}5298.\mathbf{E_{HOMO}}^2 - 24{,}26780.\mathbf{E_{LUMO}}^2 + 10{,}6816.\mathbf{\Delta E}^2 + 0{,}1999.\mu^2 + 48{,}3537.\chi^2 - 17{,}4381.\mathbf{f_{(SO)}}^2 \quad \text{(equation 2)}$$

$$\mathbf{N = 15 \qquad R = 0{,}994 \qquad R^2 = 0{,}989}$$

With MLNR was obtained significantly better correlation coefficient R = 0, 994

Figure 7 shows a very uniform distribution of the toxicity observed values depending on the experimental values and the correlation between the experimental results and calculated alter them pI_{50}. The residual values tended to zero which is why we did not graph for prediction residuals.

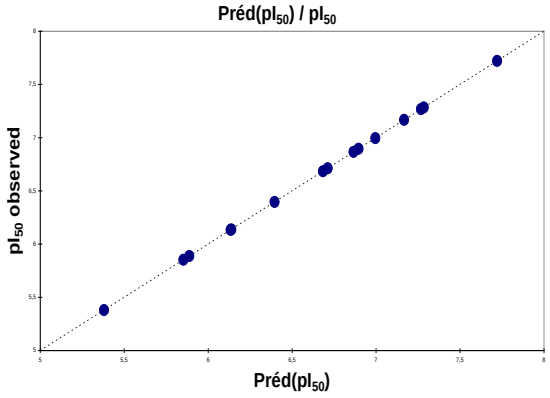

Fig. 7: Relationship between the estimated values of pI_{50} and their predictions established by MNLR

D. Partial least square regression (PLS)

To linearly correlate the molecule descriptors: the total energy (Et), .ΔE, energy E_{HOMO}, energy E_{LUMO}, the dipole moment μ, absorption maximum□χ and the factor of oscillation $f_{(SO)}$ to pI_{50}, the following equation was used:

$$\mathbf{pI_{50}} = 6{,}1319 - 2{,}5812.10^{-5}.\mathbf{Et} + 0{,}3825.\mathbf{E_{HOMO}} - 0{,}39301.\mathbf{E_{LUMO}} - 0{,}52226.\mathbf{\Delta E} + 0{,}20778.\mu\square\square + 0{,}44213.\chi\square - 6{,}6984.10^{-2}\ \mathbf{Ea} + 3{,}3369.\mathbf{f_{(SO)}} \quad \text{(equation 3)}$$

$$\mathbf{N = 15 \qquad R = 0{,}996 \qquad R^2 = 0{,}994 \qquad RMSE = 0{,}038 \qquad MCE = 0{,}001}$$

The equation 3 shows a very regular distribution of toxicity values depending on the experimental values. The obtained coefficient of correlation in equation 7 is quite interesting R = 0,996. To optimize the error standard deviation and better finish building our model, we involve in the next part artificial neural networks (ANN).

As part of this conclusion, we can say that the toxicity values obtained from past least square regression are highly correlated to that of the observed toxicity.

For our 15 compounds, the correlation between experimental and calculated toxicity one based on this model is quite significant figure 8 as indicated by statistical values.

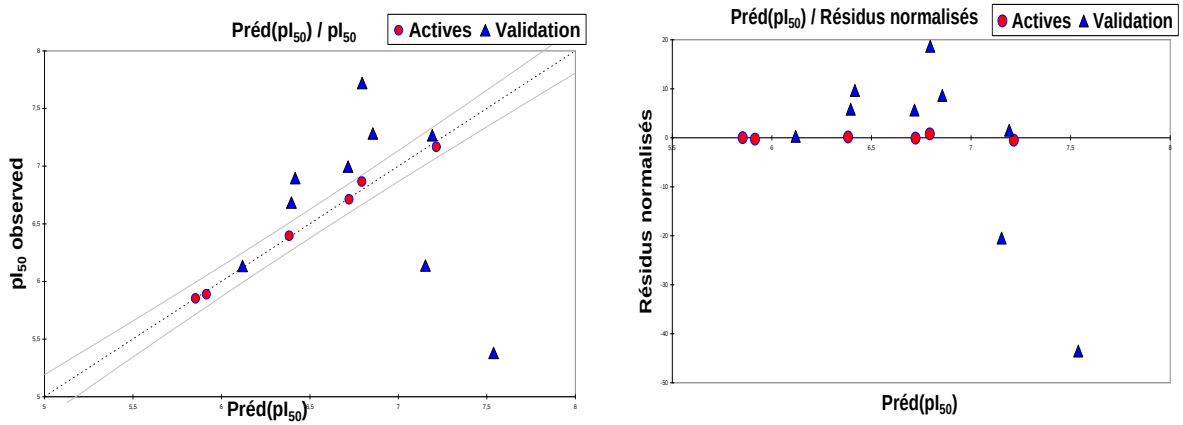

Fig. 8: Relationship between the estimated values of pI_{50}, their predictions and their residues established by partial least squares regression

E. Artificial neural networks ANN

In order to increase the probability of good characterization of studied compounds, neural networks (ANN) can be used to generate predictive models of quantitative structure–activity relationships (QSAR) between a set of molecular descriptors obtained from the MR and observed activity. The ANN calculated toxicity model was developed using the properties of several studied compounds. The correlation between ANN calculated and experimental toxicity values are very significant as illustrated in figure 8 and as indicated by R and R^2 values:

These values show that the relationship between the estimated values of pI_{50}, their residues established by artificial neural networks are illustrated in figure 9.

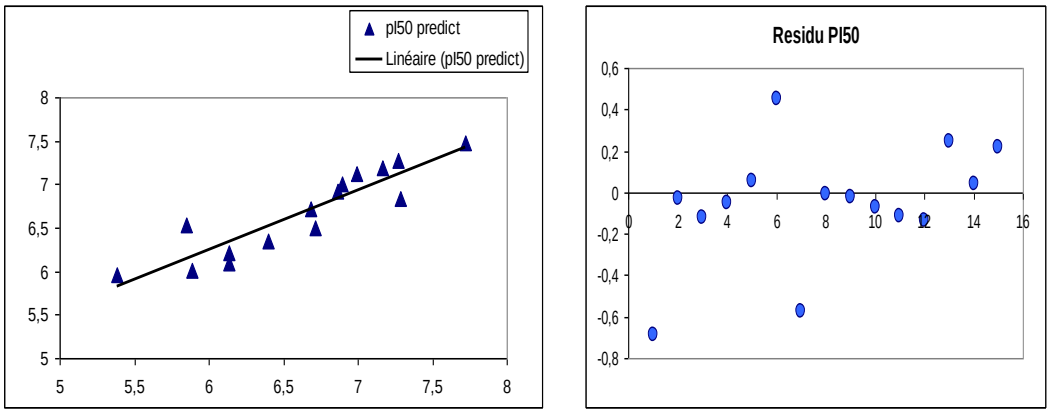

Fig. 9: Graphical representation of calculated and observed toxicity pI_{50} and their residues

The statistic of the three steps of the calculation by the ANN: training, validation and test are illustrated in table 6.

TABLE 6: VALUES OBTAINED BY ANN

RNA	Samples	MSE	R
Training	11	0,080364	0,9002
Validation	2	0,10230	0,9999
Testing	2	0,03931	0,9999

N = 15 R = 0,948 R^2 = 0,9002 MCE= 0,080364

The obtained squared correlation coefficient R value is 0.996 for this data set of 3-aryl-4-hydroxyquinolin-2-(1H)-one derivatives. It confirms that the Partial Least Square regression (PLS) results were the best to build the quantitative structure activity relationship models.

In this part, we investigated the best linear QSAR regression equations established in this study. Based on this result [30,31]. In this part, we investigated the best linear QSAR regression equations established in this study. Based on this result, a comparison of the quality of the CPA, MLR and ANN models shows that the PLS models have substantially better predictive capability because the PLS approach gives better results than MLR, MNLR and ANN. PLS was able to

establish a satisfactory relationship between the molecular descriptors and the activity of the studied compounds unlike Larif [22, 32, 33] that confirmed by ANN he found a significantly better result [34].

TABLE 7: OBSERVED VALUES AND CALCULATED VALUES OF PI_{50} ACCORDING TO DIFFERENT METHODS

pI_{50}	MLR (pI_{50})		MNLR (pI_{50})		PLS(pI_{50})		ANN (pI_{50})	
observed	**predict**	**Residu**	**predict**	**Residu**	**predict**	**Residu**	**predict**	**Residu**
5,853	6,119	-0,266	5,853	0,000	5,854	-0,001	6,5420	-0,6890
6,684	6,446	0,238	6,684	0,000	5,916	-0,028	6,7166	-0,0326
5,888	6,307	-0,419	5,888	0,000	6,793	0,073	6,0063	-0,1183
6,866	6,818	0,048	6,866	0,000	6,383	0,013	6,9131	-0,0470
6,396	6,305	0,091	6,396	0,000	7,215	-0,048	6,3413	0,0546
7,284	6,865	0,419	7,284	0,000	6,721	-0,009	6,832	0,4511
5,38	6,476	-1,096	5,380	0,000	6,396	0,288	5,955	-0,5748
7,268	7,063	0,205	7,268	0,000	6,856	0,428	7,277	-0,0090
7,167	7,054	0,113	7,167	0,000	7,539	-2,159	7,190	-0,0231
6,138	6,246	-0,108	6,138	0,000	7,192	0,076	6,206	-0,0682
6,896	7,040	-0,144	6,896	0,000	7,153	-1,015	7,009	-0,1131
6,996	7,076	-0,080	6,996	0,000	6,417	0,479	7,129	-0,1333
7,721	7,567	0,154	7,721	0,000	6,716	0,280	7,474	0,2466
6,134	6,180	-0,046	6,134	0,000	6,795	0,926	6,095	0,0394
6,712	5,820	0,892	6,712	0,000	6,120	0,014	6,496	0,2155

IV. CONCLUSION

In this work, we studied the QSAR regression to predict the toxicity of a series of 15 compounds 3-aryl-4-hydroxyquinolin-2-(1H)-one (compound structurally related to flavonoids. Comparison of key statistical terms as R or R^2 different models obtained using different statistical tools and various descriptors was shown in table 7.

The study of the quality of ANN and MLR models showed that with the PLS predictive ability was significantly better than the other methods. With the PLS approach we have established a relationship between several descriptors (E_{HOMO}, E_{LUMO},...) and the toxicity of satisfactory ways.

Finally, we can conclude that one of the descriptors studied (E_{HOMO}, E_{LUMO},...) which is sufficiently rich in chemical and electronic information to encode the structural features can be used with other topological descriptors for the development of predictive QSAR models.

ACKNOWLEDGMENT

We are grateful to the "Association Marocaine des Chimistes Théoriciens" (AMCT) for its pertinent help concerning the programs.

REFERENCES

[1] Flavin, R. et al., Future Oncol. 6, 2010, pp. 551; (b) Kridel, S. J., Lowther, W. T., Pemble, C. W. Expert Opin. Invest. Drugs, 16, pp. 1817, 2007.

[2] Kuhajda, F. P., Jenner, K., Wood, F. D., Hennigar, R. A., Jacobs, L. B., Dick, J. D., Pasternack, G. R. Proc. Natl. Acad. Sci. U.S.A., 91, pp.6379, 1994; (b) Zhou, W., Simpson, J.P., McFadden, J. M., Townsend, C. A., Medghalchi, S. M., Vadlamudi, A., Pinn, M. L., Ronnett, G. V., Kuhajda, F. P. Cancer Res., 63, pp 7330, 2003; (c) Knowles, L. M., Axelrod, F., Browne, C. D., Smith, J. W. A. J. Biol. Chem., 279, pp. 30540, 2004; (d) Pizer, E. S., Thupari, J., Han,W. F., Pinn,M. L., Chrest, F. J., Frehywot, G. L., Townsend, C. A., Kuhajda, F. P. Cancer Res., 60, pp. 213, 2000; (e) Kuhajda, F. P., Pizer, E., Li, J. N., Mani, N. S., Frehywot, G. L., Townsend, C. A. Proc. Natl. Acad. Sci. U.S.A., 97, pp. 3450, 2000; (f) De Schrijver, E., Brusselmans, K., Heyns, W., Verhoeven, G., Swinnen, J. V. Cancer Res., 63, pp. 3799, 2003.

[3] Pizer, E. S., Lax, S. F., Kuhajda, D. P., Pasternack, G.R., Kurman, R. J. Cancer, 83, pp. 528, 1997; (b) Rashid, A., Pizer, E. S., Moga, M., Milgraum, L. Z., Zahurak, M., Pasternack, G. R., Kuhajda, F. P., Hamilton, S. R. Am. J. Pathol., 150, pp. 201, 1997; (c) Milgraum, L. Z., Witters, L. A., Pasternack, G. R., Kuhajda, F. P. Clin. Cancer Res., 3, pp. 2115, 1997.

[4] Kuhajda, F. P., Jenner, K., Wood, F. D., Hennigar, R. A., Jacobs, L. B., Dick, J. D. and Pasternack, G. R., Proc. Natl. Acad. Sci. USA 91, pp. 6379-6383, 1994.

[5] Esposito, E.X., Hopfinger, A. J., Madura, J. D. Methods for applying the quantitative structure-activity relationship paradigm. Methods. Mol. Biol., 275, pp. 131-214, 2004.

[6] Perkins, R., Fang, H., Tong, W., Welsh, W.J. Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. Environ. Toxicol. Chem., 22, pp. 1666-1679, 2003.

[7] Du, Q.S., Huang, R.B., Chou, K.C. Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Curr. Protein Pept. Sci., 9, pp. 248-260, 2008.
[8] Bradbury, S.P. Quantitative structure-activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research. Toxicol. Lett., 79, pp. 229-237, 1995.
[9] Martinez-Mayorga, K., Medina-Franco, J. L. Chemoinformatics-applications in food chemistry. Adv. Food. Nutr. Res., 58, pp. 33-56, 2009.
[10] Rivkin, A., Kim, Y. R., Goulet, M.T., Bays, N., Hill, A. D., Kariv, A., Krauss, S., Ginanni, N., Strack, P.R., Kohl, N. E., Chung, C. C.,Varnerin, J. P., Goudreau, P. N., Chang, A., Tota, M. R., Munoz, Bioorg. B. Med. Chem. Lett. 16, pp. 4620-4623, 2006.
[11] Adamo, C., Barone, V. A TDDFT study of the electronic spectrum of s-tetrazine in the gas-phase and in aqueous solution. Chem. Phys. Lett., 330, pp. 152-160, 2000.
[12] Parac, M., Grimme, S. All calculations were done by GAUSSIAN 03 W software. J. Phys. Chem., A 106, pp. 6844-6850, 2003.
[13] Gaussian 03, Revision B.01, M. J. Frisch, and al., Gaussian, Inc., Pittsburgh, PA, 2003.
[14] Becke, A. D. A new mixing of Hartree–Fock and local density‐functional theories. J. Chem. Phys., 98, pp. 1372, 1993.
[15] Lee, C., Yang, W., Parr, R. G. Development of the Colle-Salvetti conelation energy formula into a functional of the electron density., Phys. Rev., B. 37, pp. 785-789, 1988.
[16] Lee, C., Yang, W., Parr, R. G. Development of the Colle-Salvetti conelation energy formula into a functional of the electron density., Phys. Rev., B. 37, pp. 785-789, 1988.
[17] Hogarh, J. N., Seike, N., Kobara, Y., Habib, A., Namd, J. J., Lee, J. S. Qilu Li, Liu, X., Jun Li, Zhang, G., Masunaga, S. Passive air monitoring of PCBs and PCNs across East Asia: A comprehensive congener evaluation for source characterization. Chemosphere, 86, pp. 718–726, 2012.
[18] Taurino, A.M., Dello Monaco, D., Capone, S., Epifani, M., Rella, R., Siciliano, P.,. Ferrara, L., Maglione, G., Basso, A., Balzarano, D. Analysis of dry salami by means of an electronic nose and correlation with microbiological methods. Sensors and Actuators B 95, pp. 123–131, 2003.
[19] Rücker, C., Rücker, G., Meringer, M.,y-Randomization and Its Variants in QSPR/QSAR. J. Chem. Inf. Model. 47, pp. 2345-2357, 2007.
[20] Nguyen N.T. Advanced Methods for Inconsistent Knowledge Management Springer- Verlag London, 2009.
[21] Demuth, H., Hugan, M., Beal M. Neural Network Toolbox. For use with MATHLAB, User Guid's, Version 9, 2011.
[22] Larif, M., Adad A., Hmamouchi, R., Taghki, A. I., Soulaymani, A., Elmidaoui, A., Bouachrine, M., Lakhlifi, T., Biological activities of triazine derivatives. Combining DFT and QSAR results, article in press in Arabian Journal of Chemistry (2013), http://dx.doi.org/10.1016/j.arabjc.2012.12.033
[23] Zupan, J., Gasteiger, J. Neural Networks for Chemistry and Drug Design: An Introduction, second ed., VCH, Weinheim, 1999.
[24] Chimizou, R., Iwamura, H., Fujita, T. Agric Food Chem. 36, pp. 1276, 1988.
[25] Turkkan, N. Génie, gènes et neurones, Revue de l'Université de Moncton, 26 (1), pp. 205-221, 1993.
[26] Lee, P. Y., Chen C. Y. J. Hazard. Mater., 165, pp. 156-161, 2009.
[27] Jing, G., Zhou, Z., Zhuo, J. Quantitative structure-activity relationship (QSAR) study of toxicity of quaternary ammonium compounds on Chlorella pyrenoidosa and Scenedesmus quadricauda. Chemosphere, 86, pp. 76-82, 2012.
[28] Jonathan, N. H., Nobuyasu, S., Yuso, K., Ahsan, H., Jae-Jak, N., Jong-Sik, L. Q. L., Xiang, L., Jun; L., Gan, Z., Shigeki, M. Passive air monitoring of PCBs and PCNs across East Asia: A comprehensive congener evaluation for source characterization. Chemosphere, 86, pp. 718–726, 2012.
[29] Elhallaoui, M., Elasri, M., Ouazzani, F., Mechaqrane, A. and Lakhlifi,T. Quantitative Structure-Activity Relationships of Noncompetitive Antagonists of the NMDA Recetor: A Study of a series of MK801 Derivative Molecules Using Statistical Methods and Neural Network. Int. J. Mol. Sci, 4, pp. 249-262, 2003.
[30] Adad, A., Hmamouchi R., Idrissi Taghki A., Abdellaoui A., Bouachrine M. and Lakhlifi T. Atmospheric half-lives of persistent organic pollutants (POPs) study combining DFT and QSPR results. Journal of Chemical and Pharmaceutical Research, 5(7), pp. 28-41, 2013.
[31] Zakarya, D., Boulaamail, A., Larfaoui, E. M. and Lakhlifi, T. QSARs for DDT-Type analogs using statistical methods and neural network. SAR and QSAR in Environmental Research, 6, pp. 183-203, 1997.
[32] Zakarya, D., Larfaoui, E. M., Boulaamail, A., Tollabi, M. and Lakhlifi, T. QSARs for a series of inhibitory anilids. Chemosphere, Vol. 36, N° 13, pp. 2809-2818, 1998.
[33] Zarrok, H., Oudda, H., Zarrouk, A., Salghi, R., Hammouti, B., Bouachrine, M. Weight Loss Measurement and Theoretical Study of New Pyridazine Compound as Corrosion Inhibitor for C38 Steel in Hydrochloric Acid Solution. Der Pharma Chemica, 3 (6), pp. 576-590, 2011.
[34] Chtita, S., Larif, M., Ghamali, M., Adad, A., Hmamouchi, R., Bouachrine, M. and Lakhlifi, T. Studies of two different cancer cell lines activities (MDAMB-231 and SK-N-SH) of imidazo[1,2-a]pyrazine derivatives by combining DFT and QSAR results. International Journal of Innovative Research in Science, Engineering and Technolog, 2 (11), pp. 6586-6601, 2013.

