Moroccan Journal of Chemistry
ISSN: 2351-812X
http://revues.imist.ma/?journal=morjchem&page=login
El-chorafi & al. / Mor. J. Chem. 5 N°4 (2017) 590-599

QSAR studies of the antiproliferative activity of heterocyclic derivatives using topological descriptors

F.Z. El-chokrafi[(a)], F. Khalil[(a)], M. Bouachrine[(b)]*

[(a)] Laboratory of Applied Chemistry, Faculty of Science and Technology, University Sidi Mohammed Ben Abdellah, Fez, Morocco.
[(b)] Equipe Matériaux, Environement & Modélisation, ESTM, University Moulay Ismail, Meknes, Morocco

** Corresponding author:*
m.bouachrine@est-umi.ac.ma
Received 25 July 2017, Revised 03 Sept 2017, Accepted 17 Sept 2017

Abstract

A QSAR study of the antiproliferative activity is applied to a set of 22 molecules using the principal component analysis (PCA), multiple linear regression (MLR) (R = 0,754), multiple nonlinear regression (MNLR) (R = 0,981) and artificial neural network (ANN) (R = 0,987) method, and finally the model was validated with the cross-validation "leave-one-out" (CV-LOO) (R = 0,601), the values of the predicted activities are in agreement with the experimental results. From the results of this study it can be said that this model gives statistically significant results and shows a good stability to the variation of data in the cross-validation leave-one-out

Keywords: *Antiproliferative, MCF-7, pGI50, QSAR, MLR, MNLR, ANN, CV.*

1. Introduction

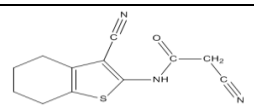
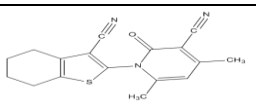
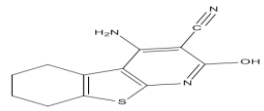
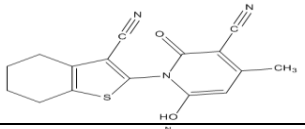
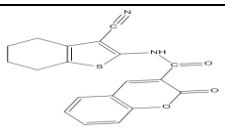
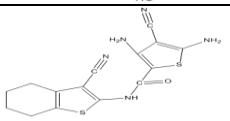
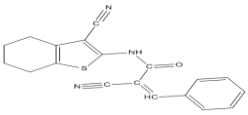
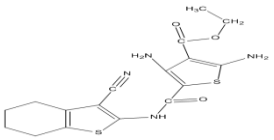
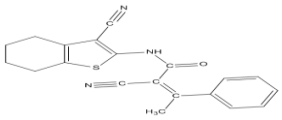
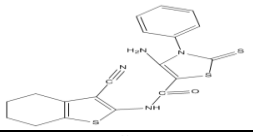
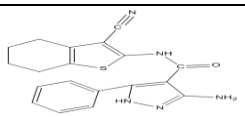
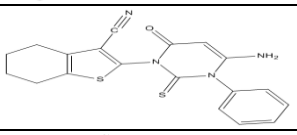
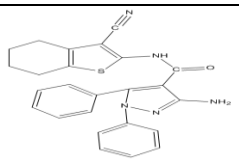
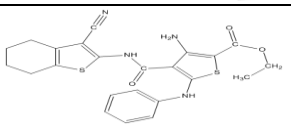
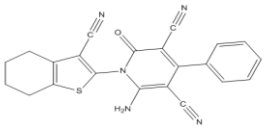
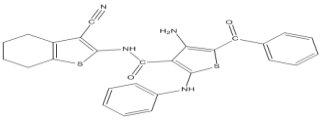
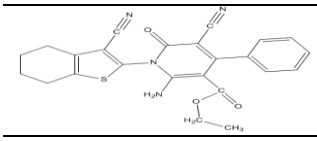
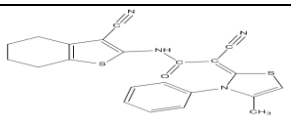
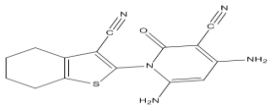
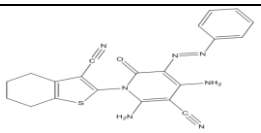
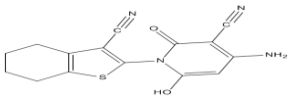
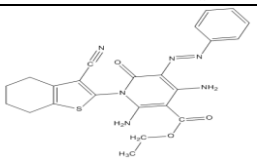
The aromatic character of heterocycles in the sulfur series contributes to their reactivity, their stability and their chemical and electronic properties. A large number of heterocyclic derivatives find in natural products [1-2], increasing application as superconductors [3-4] and optoelectronics [5-6]. The 2-cyano-N-(3-cyano-7-tetrahydrobenzo[b]thiophen-2-yl)-acetamide is used to synthesize different heterocyclic derivatives have antiproliferative [7], antimicrobial [8], antifungal [9], anti-inflammatory [10] and antioxidant [11] activities. In this article, we have been working with antiproliferative activity on MCF-7; the most widely used breast tumor cell line in breast cancer research laboratories [12]. This paper describes a study of the quantitative structure-activity relationship of the 2-cyano-N-(3-cyano-4,5,6,7-tetrahydrobenzo[b]thiophen-2-yl) that is to say to put in place a mathematical relation connecting quantitatively the molecular structure, encoded by molecular properties called descriptors, with antiproliferative activity on MCF-7 using data analysis methods. Principal component analysis (PCA) was used to classify the compounds according to their activities and to provide an estimate of the values of the relevant descriptors that govern this classification, multiple linear regression (MLR) is used to select the descriptors used as input parameters for multiple nonlinear regression (MNLR) and artificial neural network (ANN) and leave-one-out (CV-LOO) is performed to validate the proposed model.

2. Materials and methods

2.1. Experimental data

The data set consists of 22 2-cyano-N-(3-cyano-4,5,6,7-tetrahydrobenzo[b]thiophen-2-yl)-acetamide derivatives with pGI50 antiproliferative activities which have been evaluated in vitro [13] on the human tumor cell line MCF-7 (mammary adenocarcinoma), where GI50 (μM) represents the molar concentration of the compound required for 50% inhibition of the antiproliferative activity. This is to determine a quantitative structure-activity relationship between antiproliferative activity and these derivatives. Table 1 shows the substituted compounds studied and the corresponding experimental activities (pGI50).

Table 1. Antiproliferative activity observed for 2-cyano-N-(3-cyano-4,5,6,7-tetrahydrobenzo[b]thiophen-2-yl)-acetamide derivatives

No	Compound	pGI50 Observed	No	Compound	pGI50 Observed
1		-1,477	12		-1,591
2		-1,649	13		-1,342
3		-1,033	14		-1,823
4		-1,879	15		-1,342
5		-1,573	16		-1,037
6		-0,398	17		-1,629
7		-1,874	18		-1,301
8		-1,58	19		-1,072
9		-1,301	20		-1,561
10		-1,223	21		-0,301
11		-1,7	22		-1,836

2.2. Calculation of molecular descriptors

The ACD/ChemSketch [13] advanced chemical development program and ChemBioOffice 14.0 [14] were used after energy optimization for each compound using the MM2 method to calculate the following 11 descriptors:

The molecular weight (MW), the molar refractivity (MR (cm3)), the molar volume (MV (cm3)), the parachor (Pc (cm3)), the density (D (g/cm3)), the refractive index (n), the surface tension (γ (dyne/cm)), the polarizability (αe (cm3)), the lipophilic (LogP), the hydrogen bond acceptor (HBA) and the hydrogen bonding donor (HBD).

2.3. Statistical analysis

To explain the structure-activity relationship, these quantitative descriptors of 2-cyano-N-(3-cyano-4,5,6,7-tetrahydrobenzo[b]thiophen-2-yl)-acetamide derivatives are studied using statistical methods based on:

The principal components analysis (PCA), multiple linear regression (MLR) and nonlinear regression (MNLR) available on the XLSTAT program [15]. The artificial neural network (ANN) and the "leave-one-out" (CV-LOO) cross-validation are done with Matlab 7 using a program written in C language. The correlation coefficient (R), the coefficient of determination (R2), the mean squares of errors (MSE), Fisher's F-statistic (F) and Fisher's probability (Pr) are used to justify equations and select the best regression performance.

3. Results and Discussions

3.1. Data for analysis

Table 2. Descriptor values

No	MW	MR	MV	Pc	n	γ	D	Ae	LogP	HBA	HBD
1	245,3	63,52	183,7	526,3	1,607	67,3	1,33	25,18	2,518	3	1
2	245,3	66,43	163,5	508,7	1,747	93,7	1,5	26,33	3,206	4	2
3	350,391	92,42	242,5	708,9	1,687	72,9	1,44	36,63	3,869	3	1
4	333,407	92,55	252,4	724,1	1,654	67,6	1,32	36,69	4,453	3	1
5	347,433	96,95	268,5	760,4	1,641	64,3	1,29	38,43	4,629	3	1
6	363,436	99,08	251,4	767,5	1,717	86,8	1,44	39,27	3,961	5	3
7	439,532	127,57	317,6	888,7	1,735	61,2	1,38	50,57	5,86	5	2
8	397,452	107,42	270,9	828,4	1,723	87,3	1,46	42,58	3,647	5	1
9	444,506	118,6	312,5	924,7	1,683	76,6	1,42	47,01	3,772	5	1
10	311,362	81,99	206	638,6	1,727	92,3	1,51	32,5	0,843	5	2
11	312,346	79,9	199,8	625,9	1,731	96,3	1,56	31,67	1,224	5	2
12	309,386	84	231,2	659,2	1,646	66	1,33	33,3	2,921	3	0
13	311,358	80,91	212,4	636,2	1,686	80,4	1,46	32,07	3,118	4	1
14	343,427	88,97	222,8	699	1,73	96,7	1,54	35,27	2,814	5	3
15	390,48	100,15	264,6	795,3	1,681	81,6	1,47	39,7	2,939	5	3
16	412,552	112,97	268,6	836,6	1,782	94	1,53	44,78	4,503	3	2
17	380,487	105,41	253,8	786	1,769	91,9	1,49	41,79	4,819	3	1
18	466,576	124,89	327,2	967,7	1,688	76,4	1,42	49,51	5,206	5	3
19	498,619	138,6	347,4	1041,1	1,729	80,5	1,43	54,94	6,258	5	3
20	418,535	115,75	295,3	873,7	1,712	76,6	1,41	45,88	5,027	4	1
21	415,471	115,51	275,4	803,4	1,779	72,3	1,5	45,79	2,586	7	2
22	462,524	124,87	313,6	890,8	1,727	65	1,47	49,5	2,711	7	2

QSAR analysis was performed using the -log (GI50) experimental values of the selected 22 molecules that were synthesized and evaluated for their observed antiproliferative activity of 2-cyano-N-(3-cyano-4,5,6,7-tetrahydrobenzo[b]thiophen-2-yl)-acetamide. The values of the 11 chemical descriptors are shown in Table 2. The principle is to carry out first a principal component analysis, which makes it possible to eliminate highly correlated descriptors, then perform a decreasing MLR study based on the elimination of aberrant descriptors until a valid model (Pr<0,05 for all descriptors and complete model).

3.2. Principal Component Analysis (PCA)

All 11 descriptors encoding the 22 molecules were subjected to a principal component analysis (PCA). Figure 2 shows the 11 main components that were obtained.

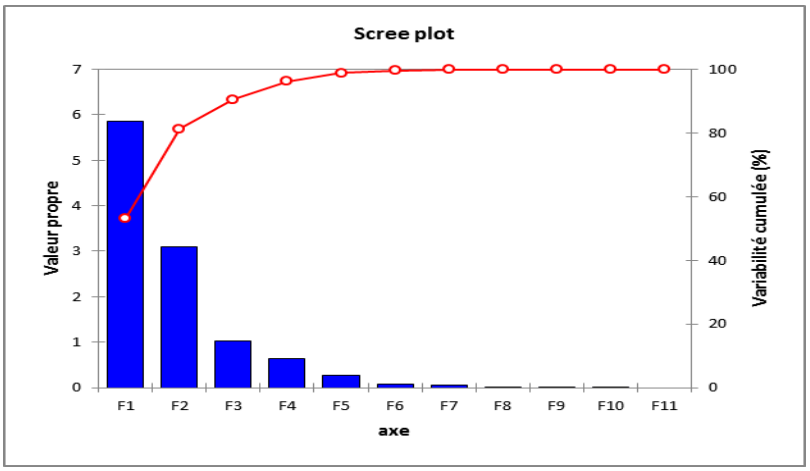


Figure 2. Principal components

With the first two axes F1 and F2 contributing 53,19% and 28,08% respectively to the total variance, the total information is estimated to be 81,27% variance. Table 3 represents the correlations between the 11 descriptors as a correlation matrix and in Figure 3 these descriptors are represented in correlation circles.

Table 3. Correlation matrix (Pearson (n)) between different descriptors obtained

Variables	MW	MR	MV	Pc	n	γ	D	αe	LogP	HBA	HBD
MW	1										
MR	0,989	1									
MV	0,967	0,974	1								
Pc	0,985	0,980	0,983	1							
n	0,353	0,369	0,151	0,258	1						
γ	-0,195	-0,247	-0,391	-0,217	0,567	1					
D	0,060	-0,006	-0,193	-0,054	0,786	0,810	1				
αe	0,989	1,000	0,974	0,980	0,369	-0,247	-0,006	1			
LogP	0,583	0,649	0,680	0,666	0,047	-0,297	-0,403	0,649	1		
HBA	0,490	0,449	0,372	0,384	0,432	0,029	0,441	0,449	-0,240	1	
HBD	0,353	0,307	0,232	0,312	0,405	0,382	0,483	0,307	0,061	0,550	1

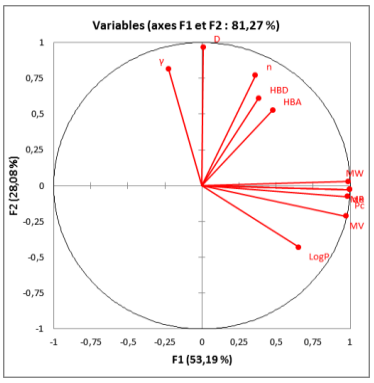


Figure 3. Correlation circle: (MR, αe) are perfectly correlated (r = 1), the two variables are redundant. MW, MR and αe are strongly correlated (r (MW, MR) = 0,989; r (MW, αe) = 0,989). MW and Pc are strongly correlated (r (MW, Pc) = 0,989. So the polarizability variable has been removed.

3.3. Multiple linear regressions (MLR)

In order to propose a mathematical model linking the descriptors and the activity and to quantitatively evaluate the physicochemical effects of the substituent on the activity of the totality of these 22 molecules, we presented the data matrix consisting of 10 variables corresponding to the training set for the multiple linear regression analysis.

Multiple linear regression allows to link the structural descriptors for the activity of each of the 22 compounds to quantitatively evaluate the effect of the substituent. The selected descriptors are: MW, MR, MV, Pc, n, γ, D, LogP, HBA and HBD. The QSAR model constructed with the multiple linear regression (MLR) method is represented by the following equation: Equation (1)

$$pGI_{50\ MLR} = -84{,}025 - 0{,}106\,(MW) - 0{,}295\,(MR) - 0{,}309\,(MV) + 0{,}19\,(Pc) + 58{,}099\,(n) - 0{,}535\,(\gamma) + 20{,}037\,(D) - 0{,}243\,(LogP) + 0{,}035\,(HBA) + 0{,}594\,(HBD)$$

N = 22 R = 0,754 R^2 = 0,569 F = 8,948 MSE = 0,024

A higher correlation coefficient, a lower mean square error indicates that the model is more reliable and the probability corresponding to the F value (Pr<0.0001) is much smaller than 0.05, this means that we would take a lower risk than 0.01% assuming that the null hypothesis is erroneous. Therefore, we can conclude that the model bring a significant amount of information. With the optimal MLR model, the predicted pIC50 MLR activity values calculated from equation (1) and the observed values are given in Table 4. The correlations of planned and observed activities are shown in Figure 4. The descriptors proposed in equation (1) by MLR were used as input parameters in the multiple nonlinear regression (MNLR) and the artificial neural network (ANN).

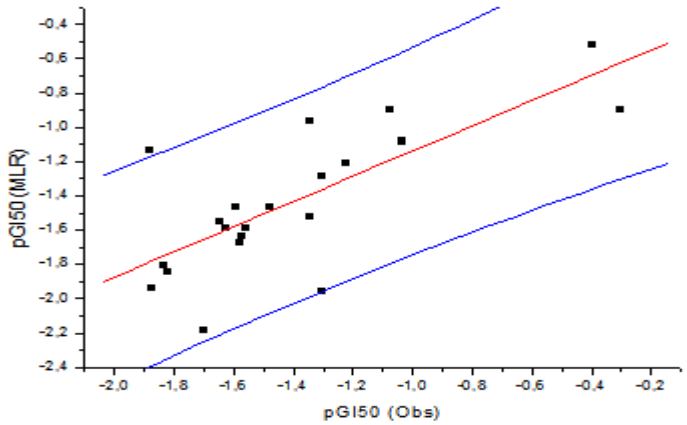


Figure 4. Graphical representation of the observed and predicted activities using the MLR

The correlation between the calculated and experimental MLR activities is significant as shown in Figure 4 and indicated by the statistical values determined, which means that the selected descriptors can form a good QSAR model.

Table 4. Observed and predicted activities (-LogGI50) according to different methods for the 2-cyano-N-(3-cyano-4,5,6,7-tetrahydrobenzo[b]thiophen-2-yl)

Molecules	pGI50 Obs	pGI50 MLR	pGI50 MNLR	pGI50 ANN	pGI50 CV
1	-1,477	-1,469	-1,434	-1,6537	-1,714
2	-1,649	-1,549	-1,614	-1,649	-1,643
3	-1,033	-1,092	-0,992	-1,0485	-1,519
4	-1,879	-1,133	-1,878	-1,6297	-1,539
5	-1,573	-1,632	-1,591	-1,6552	-1,557
6	-0,398	-0,520	-0,510	-0,4069	-1,553
7	-1,874	-1,940	-1,903	-1,8738	-1,622
8	-1,58	-1,674	-1,602	-1,5815	-1,469
9	-1,301	-1,955	-1,318	-1,3032	-1,544
10	-1,223	-1,206	-1,185	-1,223	-1,671
11	-1,7	-2,188	-1,783	-1,7026	-1,695
12	-1,591	-1,469	-1,620	-1,573	-1,556
13	-1,342	-1,519	-1,504	-1,3312	-1,49
14	-1,823	-1,847	-1,672	-1,8233	-1,627
15	-1,342	-0,960	-1,288	-1,3364	-1,439
16	-1,037	-1,080	-1,099	-1,0303	-1,037
17	-1,629	-1,593	-1,646	-1,6267	-1,644
18	-1,301	-1,288	-1,359	-1,3009	-1,312
19	-1,072	-0,897	-1,078	-1,0721	-1,302
20	-1,561	-1,588	-1,322	-1,5609	-1,576
21	-0,301	-0,896	-0,270	-0,3016	-1,081
22	-1,836	-1,803	-1,853	-1,8347	-1,595

3.4. Multiple nonlinear regressions (MNLR)

We have applied (for MNLR) to the data matrix constituted obviously from the descriptors proposed by MLR corresponding to the 22 molecules.

According to the calculations, the equation of the multiple nonlinear regression obtained is as follows: Equation (2)

$$pGI_{50\ MNLR} = -1958 + 2{,}742\,(MW) - 9{,}14\,(MR) - 7{,}776\,(MV) + 2{,}4\,(Pc) + 3202\,(n) - 4{,}666\,(\gamma) - 928{,}404\,(D) + 1{,}993\,(LogP) - 3{,}995\,(HBA) + 4{,}053\,(HBD) - 0{,}002\,(MW^2) + 0{,}024\,(MR^2) + 0{,}008\,(MV^2) - 0{,}0007\,(Pc^2) - 782{,}928\,(n^2) + 0{,}011\,(\gamma^2) + 215{,}792\,(D^2) - 0{,}617\,(LogP^2) + 0{,}268\,(HBA^2) - 0{,}717\,(HBD^2)$$

R = 0,981 R^2 = 0,962 MSE = 0,146

Table 4 give the predicted activities calculated from equation (2) and the observed values. The correlations of the predicted and observed activities are illustrated in Figure 5.

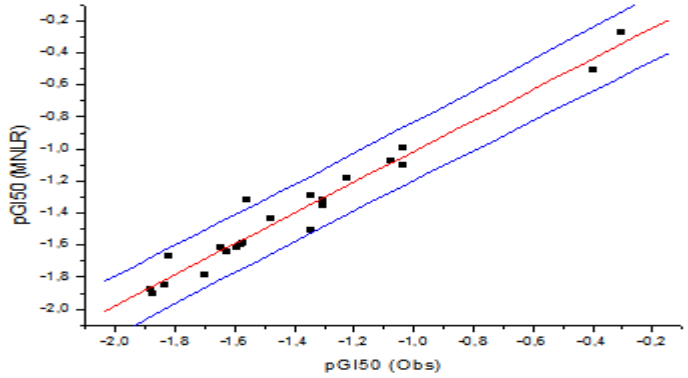


Figure 5. Graphical representation of the observed and predicted activities using the MNLR

As illustrated in Figure 5 and indicated by the values of R, R2 and MSE; the correlation between the calculated and experimental multiple nonlinear regressions activities is very significant.

3.5. Artificial neural networks (ANN)

In order to increase the probability of a good characterization of the studied compounds, artificial neural networks (ANN) can be used to generate predictive models of quantitative structure-activity relationships (QSAR) between a set of molecular descriptors obtained from MLR and observed activity. The activity model computed by the ANN was developed using the properties of several studied compounds. The values of the predicted pGI50 ANN activities calculated using ANN and the observed values are given in Table 4. Figure 6 represents the correlations of predicted and observed activities.

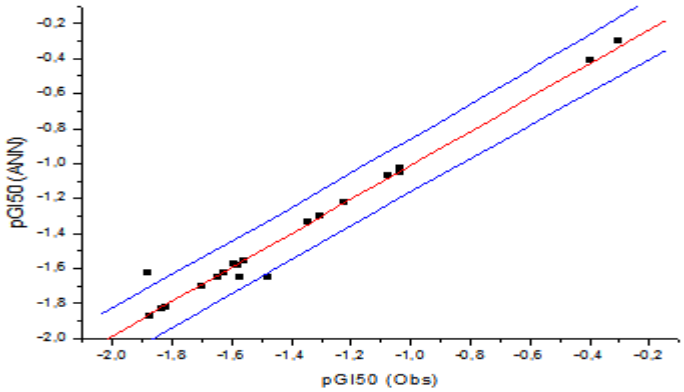


Figure 6. Graphical representation of observed and predicted activities using ANN: R = 0,987, R^2 = 0,974

As illustrated in Figure 6 and as indicated by the values R and R2, the correlation between the activities calculated by the ANN and the experimental activities is very significant. The value of the correlation coefficient (R2) obtained confirms that the result of the artificial neural network was the best for constructing the quantitative structure of activity relation models because the ANN approach gives better results than the MLR and MNLR. It is important to be able to use ANN to predict the activity of new compounds. To evaluate the predictive capacity of ANN models, the "Leave-one-out" option is an approach that is particularly well adapted to the estimation of this capacity.

3.6. Cross-validation (CV)

To test the performance of the neural network and the validity of our choice of descriptors selected by the methods used, we used the cross-validation method (CV) with the leave-one-out (LOO) procedure. In this procedure, a compound is removed from the data set; the network is formed with the remaining compounds and used to predict the rejected compound. The process is repeated in turn for each compound in the data set. The values of the predicted pGI50 CV activities calculated and the observed values are given in Table 4. The correlations of predicted and observed activities are shown in Figure 7. The correlation between the calculated predicted pGI50 CV activities and the experimental activities is significant as illustrated in Figure 7 and as indicated by the value of the correlation coefficient. The results obtained with the cross-validation show that the model proposed in this work are able to predict the activity with a high performance and that the selected descriptors are relevant.

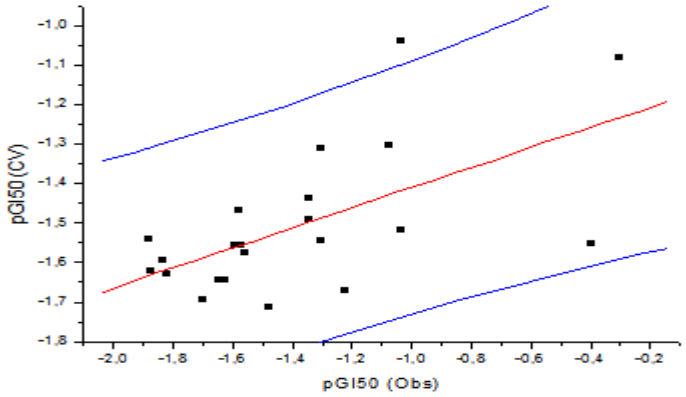


Figure 7. Graphical representation of observed and predicted activities using CV: R = 0,601, R^2 = 0,361

4. Conclusion

A multiple linear and non-linear regression and an artificial neural network were used to construct a quantitative structure-activity relationship model for 2-cyano-N-(3-cyano-4,5,6,7-tetrahydrobenzo[b]thiophen-2-yl)-acetamide. Then this model was validated by a cross validation method using the leave-one-out procedure. The artificial neural network had a significantly better predictive capacity than the other two methods, with greater predictive power. The results obtained for and show that the model proposed in this document can accurately predict the activity and that the descriptors studied sufficiently rich in chemical information to code the structural characteristics for the development of predictive QSAR model.

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

References

[1] V. Chobot, J. Vytlačilová, L. Kubicová, L. Opletal, L. Jahodář, I. Laasko, P. Vuorela. Phototoxic activity of a thiophene polyacetylene from Leuzea carthamoides. Fitoterapia (2006), 77, 194-198.

[2] Y. Tor, S.D. Valle, D. Jaramillo, S.G. Srivatsan, A. Rios, H. Weizman. Designing new isomorphic fluorescent nucleobase analogues : the thieno[3,2-d]pyrimidine core. Tetrahedron (2007), 63, 3608-3614.

[3] M. Okutan, Y. Yerli, S.E. San, F. Yilmaz, O. Günaydin, M. Durak. Dielectric properties of thiophene based conducting polymers. Syn. Metals (2007), 157, 368-373.
[4] S. Ogawa, H. Muraoka, K. Kikuta, F. Saito, R. Sato. Design of reversible multi-electron redox systems using benzochalcogenophenes containing aryl and/or ferrocenyl fragments. J. Organometal. Chem. (2007), 692, 60-69.
[5] S. Destri, U. Giovanella, A. Fazio, W. Porzio, B. Gabriele, G. Zotti. Poly(bithiophene)-co-3,4-di(methoxycarbonyl)methyl thiophene for LED. Org. Electron. (2002), 3, 149-156.
[6] M. Bouachine, O. Benaqqa, H. Toufik, M. Hamidi, J.-P. Lère-Porte, F. Serein-Spirau, A. Amine. Experimental and quantum chemical investigation of new electroluminescent material based on thiophene, phenylene and antracene. Analele Universităţii din Bucureşti (2010), 19, 35-44.
[7] L. Brault, E. Migianu, A. Néguesque, E. Battaglia, D. Bagrel, G. Kirsch. New thiophene analogues of kenpaullone: synthesis and biological evaluation in breast cancer cells. Eur. J. Med. Chem. (2005), 40, 757-763.
[8] W.W. Wardakhan, H.Z. Shams, H.E. Moustafa. Synthesis of polyfunctionally substituted thiophene, thieno [2,3-b]pyridine and thieno[2,3-d]pyrimidine derivatives. Phosph. Sulf. Silicon (2005), 180, 1815-1827.
[9] E. Pinto, M.-J.R.P. Queiroz, L.A. Vale-Silva, J.F. Oliveira, A. Begouin, J.-M. Begouin, G. Kirsch. Antifungal activity of synthetic di(hetero)arylamines based on the benzo[b]thiophene moiety. Bioorg. Med. Chem. (2008), 16, 8172-8177.
[10] G. Dannhardt, W. Kiefer, G. Krämer, S. Maehrlein, U. Nowe, B. Fiebich. The pyrrole moiety as a template for COX-1/COX-2 inhibitors. Eur. J. Med. Chem. (2000), 35, 499-510.
[11] I.C.F.R. Ferreira, M.R.P. Queiroz, M. Vilas-Boas, L. M. Estevinho, A. Begouin, G. Kirsch. Evaluation of the antioxidant properties of diarylamines in benzo[b]thiophene series by free radical scavenging activity and reducing power, Bioorg. Med. Chem. Lett. (2006), 16, 1384-1387.
[12] https://fr.m.wikipedia.org/wiki/MCF-7
[13] Z. Hoda Shams, M. Rafat Mohareb, H. Maher Helal, E. Amira Mahmoud. Novel synthesis and antitumor evaluation of polyfunctionally substituted heterocyclic compounds derived from 2-cyano-n- (3-cyano-4,5,6,7-tetrahydrobenzo[b]thiophen-2-yl)-acetamide. Molecules (2011), 16, 52-73
[14] ACD/ChemSketch Version 4.5 for Microsoft Windows User's Guide.
[15] ACD/Labs Extension for ChemBioOffice Version 14.0 for Microsoft Windows User's Guide.
[16] XLSTAT 2015 Add-in software (XLSTAT Company). www.xlstat.com.