Oriental Journal of Chemistry Vol. **25**(1), 49-56 (2009)

Application of neural network to quantitative structure anti-HIV activity relationships of flavonoid compounds

Y. BELMILOUD [1], A. KADARI[1], L. BENAHMED[2], D. CHERQUAOUI[3], D. VILLEMIN[4] and M. BRAHIMI[1]

[1]Laboratoire de Physico-Chimie Theorique et de Chimie Informatique, Faculte de Chimie,
U.S.T.H.B., BP 32 Al-alia ; Bab-Ezouar ; Alger (Algeria).
[2]Universite de Tlemecen Faculté de chimie, Tlemcen (Algerie).
[3]Departement de Chimie, Faculte des Sciences Semlalia BP 2390
Université Cadi Ayyad, Marrakech (Morocco).
[4]Ecole Nationale Supérieure d'Ingénieurs (ENSICAEN) LCMT, UMR CNRS
6507, 6 boulevard Maréchal Juin, 14050 Caen Cedex (France).

(Received: December 10, 2008; Accepted: January 15, 2009)

ABSTRACT

Artificial neural network (NN) was constructed and trained for the prediction of the anti- human immunodeficiency virus (anti-HIV) activity for 26 flavonoîd compounds based on quantitative structure-activity relationship method (QSAR). For different models, The network, inputs were selected by the stepwise multiple linear regressions technique (MLR) by using Codessa program.NN based obtained results lead to statistical results in good agreement with the literature data. They put in evidence the importance of the molecular hydrophobicity, electronegativity and atomic charges on some key atoms in modelling flavonoid compounds' behaviour by means of QSAR approach. Nonlinear NN models are shown to give better results with good predictive anti-HIV activity than linear ones.

Key words: Flavonoid, multiple linear regressions technique MLR, quantitative structure-activity relationship method QSAR, neural network NN, anti-HIV, DFT.

INTRODUCTION

Flavonoids are characterized by a common 2-phenyl-benzopyran-4-one basic structure and constitute one of the largest groups of naturally occurring compounds (Figure 1)[1]. These compounds have been reported to display a variety of biochemical properties including antioxidant[2], antimicrobial and pharmaceutical activities[3,4]. They are also used as anti-inflammatory, antiviral, antiallergic, antibiotic, and anticarcinogenic compounds[3,5]. One of the most interesting biological properties of flavonoids is their ability to inhibit human immunodeficiency virus (HIV) transcriptase and HIV replication[6].

The diversity in molecular architecture for flavonoid compounds has made possible the development of different quantitative structure-activity relationships (QSAR), allowing the identification of molecular parameters responsible for their biological and physicochemical properties[7].

To understand the chemical mechanisms associated with the biochemical effect of flavonoid, various QSAR studies have been employed to research statistical relationships between molecular structure-derived parameters and the anti-HIV properties of flavonoids[8-11].

Anti-HIV activity for 26 flavonoid compounds[12], has been yet investigated in statistical analysis, by Multiple Linear Regression (MLR) method. The best regression equation obtained by these authors was based on the following descriptors: electronegativity (χ), atomic charge on atom C7 and CLogP[12,13].

In the present work, we preceded by a MLR study of 26 flavonoid compounds by CODESSA Pro[14] and we compare our results MLR with those obtained by literature[12]. Thereafter, we will apply the Neuron Networks (NN) for a set of 26 flavonoid compounds by selecting a wide diversity of descriptors, the principal goal of the current work is:

- To provide an application of NN to the structure-anti-HIV activity relationships of flavonoid compounds.
- To compare the results obtained by the NN to those given, by multiple linear regression (MLR) and literature [12].

RESULTS AND DISCUSSION

Quantitative structure–activity relationships (QSAR) can be used to correlate structural or property "descriptors" of molecular compounds with activities. Physicochemical descriptors, which may include parameters to account for hydrophobicity, electronic properties, and steric effects, can be determined empirically or by computational methods. In the present work, several structural and physicochemical descriptors were used to characterise the studied flavonoid derivatives.

Compounds' structures were drawn using Chem-3D molecular package[15]. The molecular geometry's were optimized using DFT/RB3LYP (restricted B3LYP) quantum chemical calculations

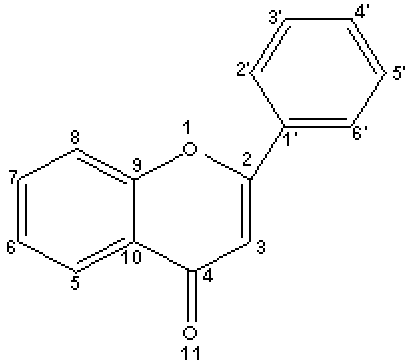

Fig. 1: Chemical structure and numbering of studied flavonoid compounds

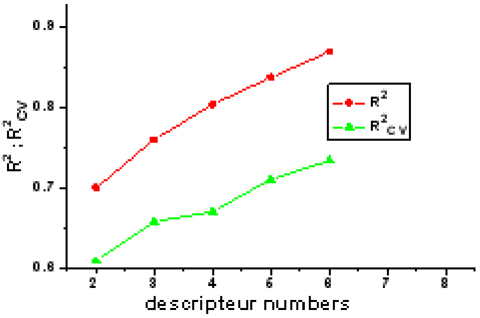

Fig. 2: R^2 ; R^2_{CV} according to the number of descriptor

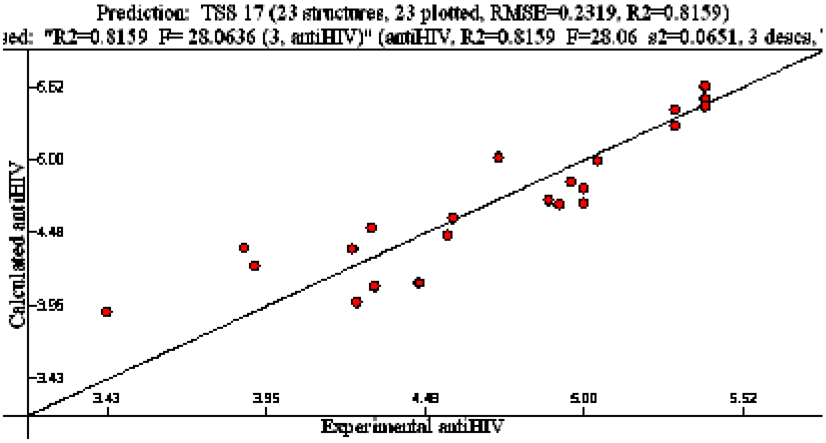

Fig. 3: Calculated Activity according to the experimental values for model 1

with the 6-31G* basis set method using Gaussian 03 [16] program package. The resulting geometry was transferred into CODESSA program[14] which can calculate more than 400 different classes of descriptors dispatched on constitutional, topological, geometrical, electrostatic, quantum chemical and thermodynamic molecular descriptor types.

CODESSA Pro program[14, 17-20] is used in the 26 flavonoid compounds MLR study

The Heuristic code Codessa program allowed the selection of three descriptors presenting the best correlation. These molecular descriptors were used as an input basis for three-layered NN (Neural Networks). Thus, descriptors and correlations are ranked according to the values of the F-test and the correlation coefficient. Starting with the top descriptor from the list, two-parameter correlations are calculated. In the following steps new descriptors are added one-by one until the pre-selected number of descriptors in the model is achieved.

The final result is a list of the 10 best models according to the values of the F-test and correlation coefficient. The quality of the correlation is tested by the coefficient regression (R^2), the Fisher ratio values (F), and the standard deviation (s^2)[19]. The stability of the models was evaluated against the cross-validated coefficient, R^2cv, which describes the stability of an obtained regression equation by focusing on the sensitivity of the model to the elimination of any single data point.

Table 1: Correlation descriptors matrix for model 1

	logP	Des1	Des2
LogP	1.0000		
Des1	-0.2151	1.0000	
Des2	-0.6981	0.3776	1.0000

Table 2: The anti-HIV activity predicted for test set by using eq.1

Structure	Experimental	Calc. Eq.1	Diff.	Calc. [12]	Diff
mol25	3.4800	3.6922	0.2122	3.860	0.380
mol26	3.4620	3.8106	0.3486	3.877	0.415

Table 3 : Correlation descriptor matrix for model 2

	logP	C7	X
logP	1.0000		
C7	0.1309	1.0000	
X	0.2144	0.0192	1.0000

For the determination of the maximum descriptor number's, ' breaking point' method was used. The variation of the standard deviation (s^2), correlation coefficient (R^2) and the cross-validation (R^2 CV) according to the number of Descriptors was shown in figure 2. The analysis of this curve shows that the break (intersection of two lines) for the coefficient of correlation is carried out for n=3. Therefore, the maximum descriptor numbers is 3 for all work.

Table 4: The anti-HIV activity predicted for test set by using eq.2

Structure	Experimental	Calc. Eq.1	Diff.	Calc. [12]	Diff
mol25	3.4800	3.8602	0.3802	3.860	0.380
mol26	3.4620	3.8774	0.4154	3.877	0.415

Multiple Linear Regression (MLR) Analysis

In a preliminary analyse, using all 24 molecules (1-24), the results indicated compound 23 to be an outlier. It is common practice in QSAR studies to omit outlier in the spirit of exploratory data analysis[12]. The following models were obtained:

Log $(1/EC_{50})$ =1.04 + 0.07ClogP + 3.23Des1 x 0.002Des 2 ...(eq.1)

$R^2 = 0.8159$ F = 28.06 $s^2 = 0.0651$ $R^2_{CV} = 0.7098$

In previous equation ClogP is logarithm of partition coefficient, Des1 is maximum of partial charge for a Hydrogen atom, Des2 is (ESP-PNSA) partial negative surface area. The squared correlation coefficient, R^2, is a measure of the fit of the regression model; R^2_{CV} is the cross-validation correlation coefficient, F the Fisher test, reflects the ration of the variance explained by the model and variance due to error in the model, high values of F-test indicate the significance of the equation.

We obtained the correlation matrix between calculated descriptors (Table.1).The CODESSA Program considers correlated variables that acquire correlation coefficient above 0.8. The correlation matrix shows that the descriptors are independent. The fit of the first model (eq.1) is shown in Fig. 3.

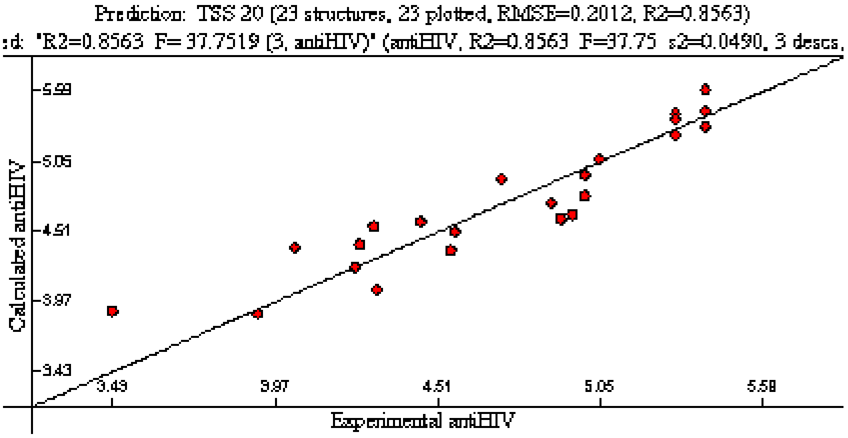

Fig. 4: Calculated Activity according to the experimental values for model 2(eq.3)

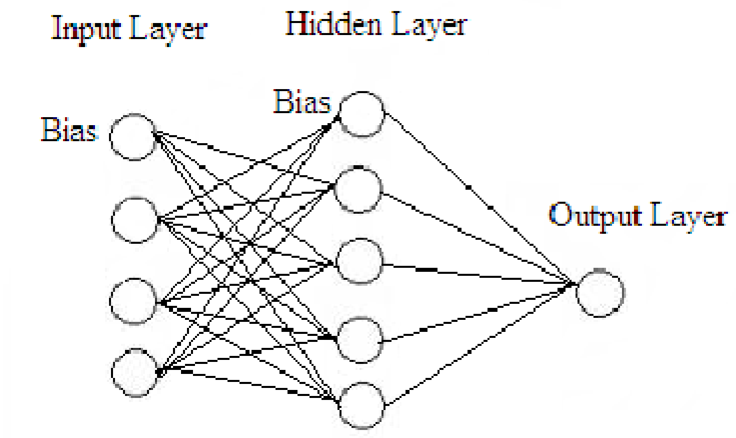

Fig. 5: The connection of different layer in this study

In order to test heuristics method capacity to selects the same descriptors suggested by literature[12] we inject their descriptor, such us electronic affinities (EA), electronegativity (÷), and charge on some key atoms such us C6, C7, C8, O11, and, Clogp in our equation. We obtain the following model:

$$\mathrm{Log}(1/EC_{50}) = 0.77 + 0.042\ \mathrm{ClogP} + 0.24\ \mathrm{C7} - 0.21\ X \qquad \text{eq.2}$$

$R^2 = 0.86$ $F=37.75$ $s^2=0.049$ $R^2_{CV}=0.7780$

The residual values indicate that the compound 2 to be an outlier. The matrix of inter correlation shows that the descriptors selected are quite independent (Table 3). In Figure.4, calculated activities according to the experimental values by eq.2 are shown.

Accordingly to the found results, its is shown that the most significant descriptors in the anti-HIV activity of the flavonoid compounds studied are ClogP is measure of hydrophobicity ; molecules with large value of ClogP have better transport through membranes, the electronegativity (X) and charge on atom 7.

For this part of the results obtained, it's can be observed that, the MLR method application to flavonoid compounds studied gave the same descriptors found by the authors of the refereed article commodity such as they were quoted, thus we found the same statistical criteria values. With regard to the prediction of the models found one observes a good prediction for model 1 (Table 2). The prediction of model 2 is the same one found by the authors (Table 4).

Table 5: Statistical Results of different NN architecture and of MLR analysis

Architecture	R²	S
3-2-1	0.944561	0.124974
3-3-1	0.946716	0.122521
3-4-1	0.955541	0.111917
MLR (Codessa)	0.856300	0.221359
Literature MLR [12]	0.860000	/

Table 6: Predictive ability of NN using leave-one-out procedure

Architecture	R² CV	Sep
3-2-1	0.837199	0.214162
3-3-1	0.699333	0.291043
3-4-1	0.662959	0.308145
MLR (Codessa)	0.7780	0.2012
Literature MLR [12]	0.80	0.45

Neural Network Analysis

NN are artificial systems emulating the function of the brain where a very high number of information processing neurons are interconnected. While there are a number of different NN models[21], the most frequently used type of NN in QSAR, and the one we shall use in this paper, is the three-layered feed-forward network. In this type of networks, the neurons are arranged in layers (an input layer, one hidden layer and an output layer), each neuron in any layer is fully connected with the neurons of a succeeding layer and no connections are between neurons belonging to the same layer. According to the supervised learning adopted in this work, the networks are taught by giving them examples of input patterns and the corresponding target outputs. Through an iterative process, the connection weights are modified until the network gives the desired results for the training set of data. A back-propagation (BP) algorithm[21] is used to minimize the error function. Neurons are connected together in layers to form the NN (Fig. 5). Typical networks have an input layer with a bias neuron, a hidden layer with another bias neuron and an output layer. The information is presented to the input layer of the network. The response of the network is coded by the output layer. The hidden layer allows the network to develop complex relationships between its input and output neurons for the training set presented. The number of neurons in the input and output layers is equal to the molecular descriptors and responses studied, respectively. The number of neurons in the hidden layer can vary, depending on the application of the network[22, 23].

In the following step, we used the three descriptors selected in model 2 for building a three-layered NN employing BP learning strategy. These three neurons constitute the input layer and describe the three variables chosen by the MLR analysis and strengthened by NN. One neuron, which encodes the anti-HIV activity, constitutes the output layer and the hidden layer contains a variable number of neurons.

Training Stage

In our case a three layered NN was used. Three neurons constitute the input layer and describe the three variables chosen by the MLR analysis and strengthened by NN. One neuron, which encodes the anti-HIV activity, constitutes the output layer and the hidden layer contains a variable number of neurons.

In this work, the number of the hidden neurons was varied from two to four in order. A bias term was added to the input and the hidden layers. The input values were normalized to [0.1 - 0.9] interval. The sigmoid function was used as the transformation function and the delta rule as the error correction formula. The weights were initialized to random values between –0.5 and +0.5 and no momentum were added. The learning rate was initially set to 1 and was gradually decreased during training. ABP algorithm implemented in C language was developed[22,23].

Three NN architectures were then trained. The optimal number of iterations required was 10000 iterations[22,23]. The results of QSAR done by these NN architectures and by MLR analysis are listed in Table 5. The quality of the fitting is estimated by the standard error of calculation (S) and by the correlation coefficient (R^2).

From Table 5 one can easily notice that all NN architectures trained show high fitting ability. The high correlation coefficients given by the trained NN architectures indicate that the log(1/EC_{50}) activity is significantly correlated with the three variables adopted in this work.

It is noteworthy that the results of the NN are significantly better than those obtained by MLR analysis. Usually, for QSAR containing non-linear relationships NN put up better performances than MLR [24]. This provides evidence for the non-linearity of the relationship between the structural features of flavonoid compounds and anti-HIV activity.

Prediction stage

It is obvious that medicinal chemists take a keen interest in the design of new drugs. What is needed is a system that is able to provide reasonable predictions for the compounds that are previously unknown. Indeed, one of the most important attributes of NN is their ability to generalize[25] that is their ability to make predictions on new data with accuracy similar to that with the training set. Besides, NN are known for their ability to model a wide set of functions without knowing the analytic forms in advance. After training, The NN is initiated to recognize the relationship between input and output data and creates an internal model as a governing data process. The NN can then use this internal model to make predictions for new inputs.

After determining the range of hidden neurons giving a good computation, the most important predictive aspect of NNs was studied: the prediction of the anti-HIV activity of new molecules. To determine that predictive aspect, leave-one – out procedure[26] has been used. In this procedure one compound is removed from the data set, the network is trained with the remaining compounds and used to predict the discarded compound. The process is repeated in turn for each compound in the data set. The analysis of predictive ability was carried out in terms of both predictive r^2 and standard error of prediction SEP.

The results obtained are presented in table 6. The 3-2-1 NN architecture exhibits a good predictive performance. Besides, its fitting ability seems satisfying. The ability to generalize being the most important criterion, therefore the 3-2-1 NN architecture is well adapted to relate anti-HIV-1 activity of flavonoid compounds to their structural requirements, and then it is adopted for the analyses that will follow.

CONCLUSION

This article is a complete work of molecular modelling, giving all the stages of the application of the MLR and NN in anti-VIH fields QSAR of the flavonoid compounds. Indeed, the application of these methods on the 26 flavonoid compounds gives:

´ The same results found in the literature concerning the importance of hydrophobic parameter, the electronegativity (X) and charge on atom 7.

´ Comparison of the correlation models obtained by MLR and NN, it can be seen that the performance of NN is better than of the results obtained by the literature.

REFERENCES

1. E. de Rijke, P. Out, W.M.A. Niessen, F. Ariese, C. Gooijer, U.A.Th. Brinkman, *J. Chromatogr.* A, **1112**: 31 (2006).
2. C.A. Rice-Evans, N.J. Miller, G. Paganga, *Trends Plant Sci*. **2**: 152 (1997).
3. W. Heller,G. Forkmann, in: J.B. Harborne (Ed.), The Flavonoids, Chapman & Hall, London, 399 (1988).
4. E. Wollenweber, in: V. Cody, E. Middleton Jr., J.B. Harborne, A. Beretz (Eds.), Plant Flavonoids in Biology and Medicine, II. Biochemical, Cellular and Medicinal Properties, Liss, New York, 45 (1988).
5. K. Janssen, R.P. Mensink, F.J. Cox, J.L. Harryvan, R. Hovenier, P.C. Hollman, M.B. Katan, *Am. J. Clin. Nutr.* **67**: 255 (1998).
6. Jaime Souza, Jr, Regina Helena de Almeida Santos, Marcia Miguel Castro Ferreira, Fabio Alberto Molfetta, Ademir Joao Camargo, Kathia Maria Honorio, Alberico Borges Ferreira da Silva. *European Journal of Medicinal Chemistry* **38**: 929-938. A quantum chemical and statistical study of flavonoid compounds (flavones) with anti-HIV activity (2003).
7. Huang, X.; Liu, T.; Gu, J.; Luo, X.; Ji, R.; Cao, Y.; Xue, H.; Wong, J. T.; Wong, B. L.; Pei, G.; Jiang, H.; Chen, K. *J. Med. Chem.*, **44**: 1883-1891 (2001).
8. Novic, M.; Nikolovska, Z.; Solmajer, T. J. *Chem. Inf. Comput. Sci.* **37**: 990-998 (1997).
9. Huang, X.; Liu, T.; Gu, J.; Luo, X.; Ji, R.; Cao, Y.; Xue, H.; Wong, J. T.; Wong, B. L.; Pei, G.; Jiang, H.; Chen, *K. J. Med. Chem.,* **44**: 1883-1891 (2001).
10. Marder, M.; Estiu, G.; Blanch, L. B.; Viola, H.; Wasowski, C.; Medina, J. H.; Paladini, A. C. Bioorg. *Med. Chem.*, **9**: 323-335 (2001).
11. Luco, J. M.; Yamin, L. J.; Ferretti, H. F. *J. Pharm. Sci.,* **84**: 903-908 (1995).
12. J. Lameira, C.N. Alves, V. Moliner, E. Silla. *European Journal of Medicinal Chemistry* **41**: 616-623 (2006).
13. B.R. Kowalski, Chemometrics “Mathematics and statistics in chemistry”, D. Reidel Publ Comp, Dordrecht, (1984).
14. A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA: Reference Manual, University of Florida, Gainesville, FL, (1994).
15. CS Chem 3D Ultra Molecular Modeing and Analyses, Cambridge soft, 2001.
16. Gaussian 98, Revision A.9, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople, Gaussian, Inc., Pittsburgh PA, (1998).
17. A.G. Maldonado, J.P. Doucet, M. Petitjean, B.T. Fan, Mol. Divers., in press
18. M. Oblak, M. Randic, T. Solmajer, *J. Chem.*

Inf. Comput. Sci. **40**: 994 (2000).

19. F. Luan *Computational Materials Science* **37**: 454-461 (2006).
20. A.R. Katritzky, R. Petrukhin, R. Jain, M. Karelson, *J. Chem. Inf. Comput. Sci.* **41**: 1521 (2001).
21. Zupan, J. Gasteiger, J. Neural Networks in *Chemistry and Drug Design*, Wiley-VCH, N.Y., (1999).
22. Douali, L., Villemin, D., Cherqaoui, D., *Curr. Pharm. Des.* **9**: 1817-1826 (2003).
23. Latifa Douali, Didier Villemin and Driss Cherqaoui, *Int. J. Mol. Sci.* **5,** 48-55 (2004).
24. Gasteiger, J.; Zupan, J.; Neural Networks in Chemistry. *Angew. Chem. Int. Ed. Engl.* **32**, 503-527 (1993).
25. Bishop, C. M. Neural Networks and their *Applications. Rev. Sci. Instrum.*, **65**: 1803-1832 (1994).
26. Zakarya, D.; Cherqaoui, D.; Esseffar, M.; Villemin, D.; Cense, *J. M. J. Phys. Org. Chem.,* **10**: 612-622 (1997).