Neural Comput & Applic (2013) 23:577–582
DOI 10.1007/s00521-013-1404-0

Generation of artificial neural networks models in anticancer study

Inês J. Sousa · José M. Padrón · Miguel X. Fernandes

Received: 31 January 2013 / Accepted: 2 April 2013 / Published online: 23 April 2013

Abstract Artificial neural networks (ANNs) have several applications; one of them is the prediction of biological activity. Here, ANNs were applied to a set of 32 compounds with anticancer activity assayed experimentally against two cancer cell lines (A2780 and T-47D). Using training and test sets, the obtained correlation coefficients between experimental and calculated values of activity, for A2780, were 0.804 and 0.829, respectively, and for T-47D, we got 0.820 for the training set and 0.927 for the test set. Comparing multiple linear regression and ANN models, the latter were better suited in establishing relationships between compounds' structure and their anticancer activity.

Keywords Backpropagation algorithm · Correlation coefficients · Heuristics · Learning algorithms · Machine learning · Neural network models · Nonlinear models · Prediction methods · Radial base function network

1 Introduction

Quantitative structure–activity relationship (QSAR) is the broad designation of several statistical methods used to establish the relation between parameters that describe molecular structure (molecular descriptors) and a certain biological activity of interest shown by a series of compounds under study [1]. The ultimate goal is to establish a statistical model that will allow the prediction of activity of novel compounds before laborious and expensive synthesis and experimental testing [2]. Relation between descriptors (independent variables) and activity (dependent variable) can be retrieved using linear and nonlinear methods, and accepted models must allow an easy translation of its parameters into synthetic chemical features. Multiple linear regression (MLR) has been used to establish QSAR for decades but the use of artificial neural networks (ANNs) still bears the charm of novelty [3]. For QSAR models to work properly, all the compounds must belong to a congeneric series (synthesized from a common chemical scaffold) and share the same mode of action to trigger the biological activity [2, 4]. Linear methods, like MLR, have the advantage of being easier to implement and interpret. Nonlinear methods, like ANN, have the advantage of producing, in general, better relations since biological response is a very complex phenomenon, sometimes not very well described by linear relations [5]. The downside of nonlinear models is the difficulty of their interpretation [6]. Irrespective of the method (linear or nonlinear), the statistical model is created using a training set of compounds and the predictive ability of the model is evaluated using a test set (compounds from the test set were not used to create the statistical model) [7]. In this type of applications, QSAR models are accepted if the correlation coefficient between experimental and calculated values of biological activity is better than 0.8 for the training set, and if the correlation coefficient between experimental and calculated values of biological activity is better than 0.6 for the test set [8]. Here, we show the application of ANN to create QSAR models for a series of compounds with experimentally determined anticancer activity.

I. J. Sousa · M. X. Fernandes (✉)
Centro de Química da Madeira, Centro de Competência de Ciências Exatas e de Engenharia, Universidade da Madeira, Campus da Penteada, 9000-390 Funchal, Portugal
e-mail: mxf@uma.pt

J. M. Padrón
BioLab, Instituto Universitario de Bio-Orgánica "Antonio González", Universidad de La Laguna, C/Astrofísico Francisco Sánchez 2, 38206 La Laguna, Spain

2 Methodology

2.1 Dataset for analysis

A set of 32 compounds (Fig. 1) derived from natural products, which showed anticancer activity in cellular assays, was used in this study. Their activity was assessed against two cancer cell lines: human ovarian cancer cell line (A2780) and mammary carcinoma cell line (T-47D). The activity is expressed in IC_{50} values (concentration of compound needed to inhibit 50 % of the cancer cells in a sample). The number of compounds used to establish the correlations for A2780 and T-47D cell lines was 23 and 22, respectively.

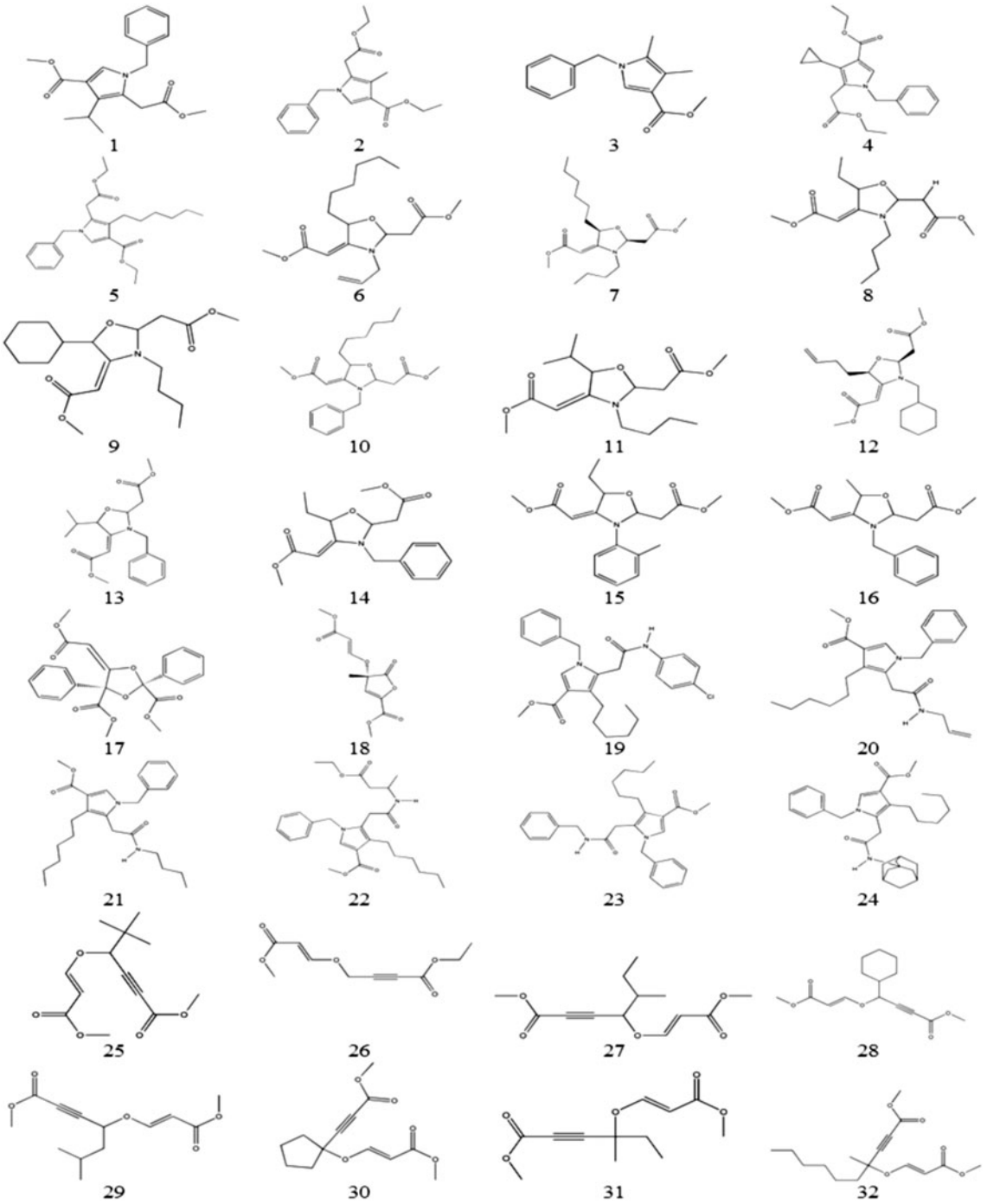

Fig. 1 Structures of anticancer compounds under study

2.2 Computational details

The optimization of compounds' structures was performed using molecular mechanics (MM+) force field included in HyperChem Professional 7.51 [9]. MOPAC [10] included in Vega ZZ 2.2.0 [11] was applied to calculate thermodynamic and quantum chemical descriptors using the following keywords: "FORCE PRECISE THERMO ROT = X" and "VECTORS BONDS PI POLAR PRECISE ENPART," respectively. Additional descriptors, as number of H-bond donors and acceptors, and logP, were calculated using E-DRAGON [12].

These descriptors along with MOPAC output files and HyperChem structure files were used as input files in CODESSA program [13] to calculate additional molecular descriptors. CODESSA calculates five classes of molecular descriptors: constitutional; topological; geometrical; electrostatic; and quantum chemical. Constitutional descriptors are associated with the number of atoms and bonds in each molecule. Topological descriptors include valence and nonvalence molecular connectivity indices calculated from the hydrogen suppressed formula of molecules, and encode information about size, composition, and the degree of branching of a molecule. The quantum chemical descriptors provide information about binding and formation energies, partial atom charge, dipole moment, and molecular orbital energy levels. A total number of 263 descriptors were calculated for each molecular structure.

2.2.1 *Heuristic method*

The heuristic method, implemented by CODESSA, was used to perform molecular descriptors selection based on their individual correlation with the biological activity. This method performs the elimination of descriptors discarding those that satisfy at least one of the following conditions: (a) the descriptor value is not available for every structure; (b) the descriptor has a constant value for all structures. After this elimination, the one-parameter correlation equations for each descriptor are calculated. To reduce even further the number of descriptors in the initial set, the following criteria are applied and descriptors are eliminated if (a) the F test's value for the one-parameter correlation with the descriptor is below 1.0; (b) the squared correlation coefficient of the one-parameter equation is less than $R^2_{\min}$ (0.01); (c) the parameter's t value is less than t_1(0.1); (d) the descriptor is highly intercorrelated (above r_{full}, where r_{full} is a user-specified value) with another descriptor with a higher squared correlation coefficient in the one-parameter equations based on these descriptors. All the remaining descriptors are then listed in decreasing order of their regression coefficients for the corresponding one-parameter correlation equations.

2.3 Training and test set selection

Each set of compounds was divided into five subsets according to their biological activity range. The compounds for training and test sets were selected randomly from within each subset in order to ensure the diversity of training set and to guarantee that test set compounds were representative of the dataset. This selection was performed keeping in view the training/test set ratio of 4:1. Based on these rules, the test sets were built with 5 compounds for the two cancer cell lines under study, and training sets had 18 and 17 compounds for A2780 and T-47D, respectively. In Tables 1 and 2, we present training and test sets and the biological activity values, predicted and experimental.

2.4 Artificial neural networks application

Statistica 7 [14] was used to perform the ANN methodology. The biological activity values (IC_{50}) and the values of molecular descriptors most correlated with the property were used to develop the ANN models. In Statistica 7 program, a quick regression method was applied using the intelligent problem solver to perform the analysis.

Table 1 Experimental and predicted activity values for A2780 cancer cell line

Comp. no.	Exp.	MLR	ANN
5[a]	4.918	4.396	4.552
6[a]	4.580	4.446	4.567
7	4.590	4.493	4.475
9	4.720	4.444	4.499
11	4.440	4.462	4.514
12	4.580	4.579	4.667
13[a]	4.410	4.512	4.659
17	4.444	4.765	4.822
18	4.982	4.975	4.818
19	5.147	4.913	4.675
20	4.260	4.471	4.538
21	4.247	4.436	4.528
22	4.806	4.806	4.724
23	4.277	4.495	4.558
24	4.880	4.473	4.555
25	5.551	5.838	5.815
26	5.565	5.568	5.797
27[a]	7.984	5.932	5.815
28	5.707	5.770	5.822
29	6.458	5.988	5.816
30	5.997	5.917	5.812
31[a]	6.356	5.944	5.809
32	5.609	5.866	5.821

[a] Compounds from test set

Table 2 Experimental and predicted activity values for T-47D cancer cell line

Comp. no.	Exp.	MLR	ANN
1	4.267	4.383	4.128
2	4.159	4.034	4.223
3	4.144	3.914	4.134
4[a]	4.278	4.422	4.239
5[a]	4.625	4.102	4.249
6[a]	4.189	4.048	4.114
7	4.204	4.079	4.115
8	4.000	4.149	4.115
9	4.316	4.192	4.120
10	4.235	4.142	4.233
11	4.096	4.114	4.117
12	4.072	4.192	4.206
13	4.056	4.074	4.115
14[a]	4.000	4.149	4.117
15	4.407	4.334	4.332
16	4.000	4.327	4.115
25	4.678	5.062	5.322
27	5.272	5.326	5.322
29	5.082	5.310	5.322
30[a]	5.599	5.719	5.322
31	5.691	5.300	5.322
32	5.886	5.632	5.322

[a] Compounds from test set

The biological activity was selected as continuous output, the molecular descriptors are the continuous input, the subset variable corresponds to the designation of subset of each compound, and may be considered as either a training set or a test set. Three types of neural networks were selected to be tested, linear, radial basis function (RBF) and multilayer perceptrons. Relatively to the network complexity, the number of neurons in the hidden layer was selected to vary between 1 and 3 to RBF, and 1 and 3 to MLP method. Some algorithms as K-means, K-nearest neighbor, pseudo-inverse, back-propagation and conjugate gradient descendent were used to train the artificial neural networks.

2.5 Multiple linear regression

MLR methodology was applied to the same training and test sets and using IC_{50} values for A2780 and T-47D cancer cell lines. This methodology was implemented using heuristic method which establishes correlations between biological activity and molecular descriptors. The obtained correlations are evaluated based on the values of statistical criteria (F and t test, R^2 and RMSE) and selected the model with best predictive ability. These linear models were compared with nonlinear models obtained previously.

3 Results and discussion

3.1 Human ovarian cancer cell line (A2780)

The application of artificial neural network to the biological activity (IC_{50}) values and the molecular descriptors resulted in a nonlinear model obtained through multilayer perceptrons method. The resulting neural network shows a structure of 3-2-1, as represented in Fig. 2.

The molecular descriptors involved in this model are as follows: number of triple bonds; relative negative charge; and average information content (order 0).

The number of triple bonds belongs to constitutional molecular descriptors group, and it is related to compounds' reactivity [15]. The quantum chemical molecular descriptor involved in this model is the relative negative charge which represents the charge distribution and describes the electrostatic interactions of molecules [16]. The average information content (order 0) is a topological molecular descriptor that is associated with the dispersion interactions of molecules and describes their size, branching, and composition [1].

The analysis of artificial neural networks does not allow an easy interpretation of molecular descriptors contribution to the biological activity, but based on the ratio between the performance of neural network before and after the elimination of each descriptor (sensibility analysis) is possible to determine the significance of molecular descriptors. The application of sensibility analysis to the obtained neural network allowed us to identify the number of triple bonds as the most significant descriptor for this model, followed by the relative negative charge and the average information content (order 0).

In Fig. 3, we show the correlation between predicted and experimental activity values obtained through artificial neural networks. The resulting correlation coefficients were 0.804 and 0.829 for training and test sets, respectively.

3.2 Mammary carcinoma cell line (T-47D)

In order to establish a nonlinear relationship between biological activity (IC_{50}) and molecular descriptor for this

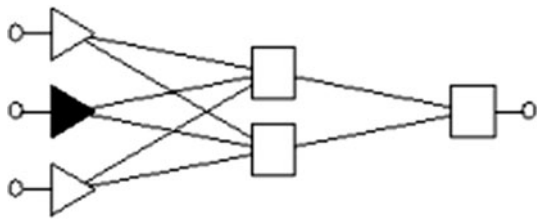

Fig. 2 Neural network (3-2-1) design obtained for A2780 cancer cell line

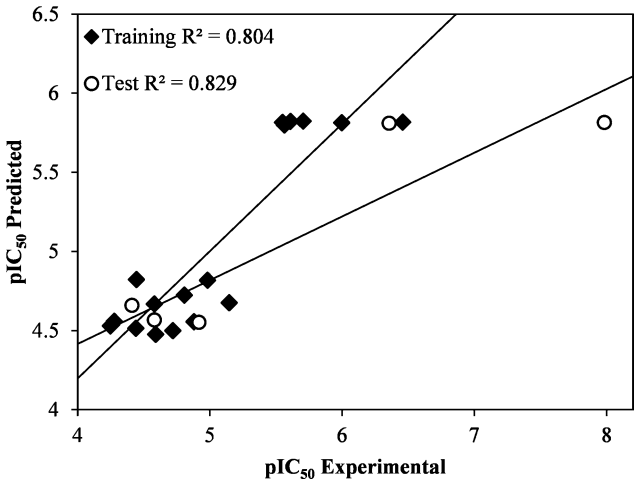

Fig. 3 Correlation between experimental and calculated activity values for A2780 cancer cell line

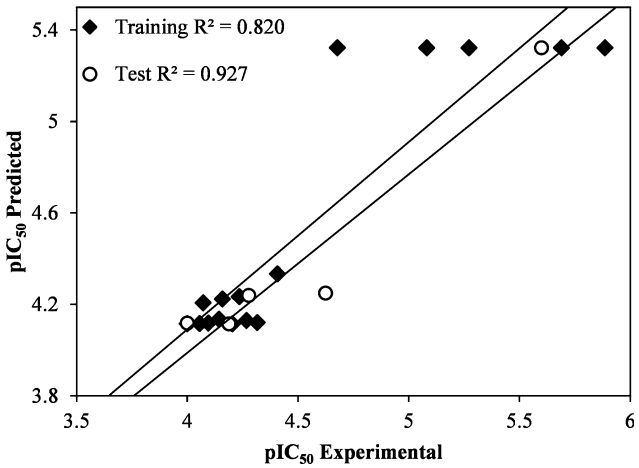

Fig. 5 Correlation between experimental and calculated activity values for T-47D cancer cell line

cancer cell line, artificial neural networks were applied. From this was obtained a neural network with the structure 2-3-1 through RBF method, as represented in Fig. 4.

In this model, two molecular descriptors were used: the number of N atoms and the area-weighted surface charge of hydrogen-bonding donor atoms. Performing the sensitivity analysis, we verified that area-weighted surface charge of hydrogen-bonding donor atoms is the most significant descriptor for this model. This is an electrostatic descriptor which is related to the hydrogen-bonding acceptor properties of the molecules [16]. The number of N atoms belongs to constitutional descriptors group, and it is related to the capability of a molecule to form hydrogen bonds [17].

The correlation between predicted and experimental activity values for training and test sets is shown in Fig. 5. For training set, the correlation coefficient obtained was 0.820 and for test set was 0.927.

3.3 Linear and nonlinear models comparison

In order to determine the best method, linear or nonlinear, to establish the relationship between molecular structure and activity, the linear methodology was applied to the same compounds set and using the experimental biological activity (IC_{50}) values obtained for A2780 and T-47D cancer cell lines.

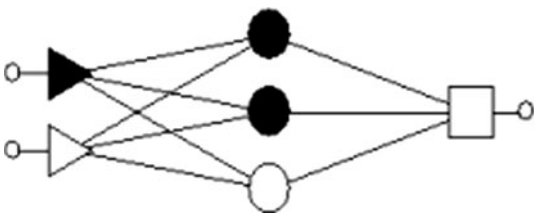

Fig. 4 Neural network (2-3-1) design obtained for T-47D cancer cell line

Table 3 Correlation coefficients (R^2) values obtained for A2780 and T-47D cancer cell lines using linear and nonlinear methods

Methods	A2780		T-47D	
	Training	Test	Training	Test
MLR	0.875	0.825	0.882	0.829
ANN	0.804	0.829	0.820	0.927

In Table 3, we show the correlation coefficients (R^2) for training and test sets using linear and nonlinear methods for the two cancer cell lines under study. Comparing the results, we can say that when using nonlinear methodology, the correlation coefficients obtained are, in general, slightly better.

4 Conclusions

ANN methodology was used to obtain nonlinear models that describe the relationship between the structure and the anticancer activity of compounds. MLP method proved to be more suitable to establish this relationship using anticancer activity values for A2780 cancer cell line and RBF method to T-47D cancer cell line. The obtained models showed high prediction ability and provided information about the structural features that influence the activity. The application of a linear method to the system under study allowed the comparison between linear and nonlinear models. In this work, the nonlinear models showed to be better suited in describing the relationship between molecular descriptors and anticancer activity for A2780 and T-47D cancer cell lines.

Acknowledgments This research was supported by BIOPHARMAC Project (BIOPHARMAC—MAC/1/C104) and Project PEst-OE/QUI/UI0674/2011.

References

1. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, Hu ZD, Fan BT (2004) Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. J Chem Inf Comput Sci 44:1257–1266
2. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design, 2nd edn. Wiley-VCH, Weinheim
3. Douali L, Villemin D, Cherqaoui D (2003) Neural networks: accurate nonlinear QSAR model for HEPT derivatives. J Inf Comput Sci 43:1200–1207
4. Gramatica P, Vighi M, Consolano F, Todeschini R, Finizio A, Faust M (2001) QSAR approach for the selection of congeneric compounds with a similar toxicological mode of action. Chemosphere 42:873–883
5. Guha R, Jurs PC (2005) Interpreting computational neural network QSAR models: a measure of descriptor importance. J Chem Inf Model 45:800–806
6. Livingstone DJ, Manallack DT (2003) Neural networks in 3D QSAR. QSAR Comb Sci 22:510–518
7. Winkler DA (2002) The role of quantitative structure—activity relationships (QSAR) in biomolecular discovery. Brief Bioinform 3:73–76
8. Hansch C, Verma RP (2009) A QSAR study for the cytotoxic activities of taxoids against macrophage (MΦ)-like cells. Eur J Med Chem 44:274–279
9. HyperChem(TM) Professional 7.51, Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA
10. Stewart J (1993) MOPAC manual, 7th edn, Fujitsu Limited, Tokyo
11. Pedretti A, Villa L, Vistoli G (2004) VEGA—an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming. J Comput-Aided Mol Des 18:167–173
12. Tetko I, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin V, Radchenko E, Zefirov N, Makarenko A, Tanchuk V, Prokopenko V (2005) Virtual computational chemistry laboratory—design and description. J Comput-Aided Mol Des 19:453–463
13. Katritzky AR, Lobanov VS, Karelson M (1994) CODESSA: reference manual. Version 2; University of Florida
14. StatSoft, Inc. (2007) Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/stathome.html
15. Katritzky AR, Petrukhin R, Jain R, Karelson M (2001) QSPR analysis of flash points. J Chem Inf Comput Sci 41:1521–1530
16. Xue CX, Zhang RS, Liu HX, Yao XJ, Liu MC, Hu ZD, Fan BT (2004) QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. J Chem Inf Comput Sci 44:1693–1700
17. Papa E, Villa F, Gramatica P (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (Fathead Minnow). J Chem Inf Model 45:1256–1266