Journal of International Scientific Publication:
Materials, Methods & Technologies, Volume 3

NEURAL NETWORKS APPLICATION FOR PREDICTION OF FLEXIBLE ORGANIC COMPOUNDS PROPERTIES

Valery V. Burlyaev, Elena V. Burlyaeva, Maxim E. Berlin

Department of Information Technology, Moscow State Academy of Fine Chemical Technology,

117571, Russia, Moscow, Prospekt Vernadskogo, 86, e-mail: compchem@mitht.rssi.ru

Abstract

The computer prediction of organic compound properties makes pharmacological investigations more cheap and soon. The prognosis is based on relations between molecular structures and properties of a series of known compounds. We use artificial neural network to model these relations and provide the technique for prediction of organic compounds properties based on ANN. We provide the regulation algorithm to improve the predictive power of the networks.

Key words: *drug design, artificial neural networks, function modeling*

1. INTRODUCTION

The traditional way to get a new drug is to synthesize and to analyze pharmacological and toxicological properties of the set of chemical compounds. Usually such compounds have the same "nucleus" and the different proxies. The investigations in this field are very expensive. We need the means for the prediction of the pharmacological properties of the new compound. The prognosis is based on casual relations between molecular structures and quantitatively measured properties of a series of known compounds. Discovering of such relations is called quantitative structure-activity research (QSAR). Unfortunately, there are no simple models revealing correlation between molecule's descriptors and its properties. Artificial Neural Networks (ANN) are modern non-linear statistical data modeling tools [1]. They are often used to model complex relationships between inputs and outputs. So we provide the technique for prediction of organic compounds properties based on ANN. We use IDEF0 (Integration Definition for Function Modeling) to describe the technique stages.

2. THE TECHNIQUE FOR THE PREDICTION OF THE FLEXIBLE ORGANIC COMPOUNDS PROPERTIES

2.1 Initial data and results

The data set consists of the descriptions of the compounds with known activity. Each compound is represented by single experimentally measured activity value and many calculated values that describe the molecule's structure. All of the data are numerical.

As was found, it is necessary to use quantum-chemical descriptors to describe the three-dimensional structure of the molecule. But the molecule is not 'fixed' in a certain 3D form during the interaction's course. So we have to consider multiple conformers of one molecule, and the fuzziness appears. Mathematically, this fuzziness means that one value of experimentally measured chemical's activity corresponds with many values of descriptors for each conformer of the same molecule. To resolve this fuzziness, we must select a single conformer. The most energetically stable conformer can be considered as such a conformer.

We choose MOPAC [2] as the program for molecular structure optimization because it is widely known and open source program. The program implements molecular mechanic and semi-empirical methods of total molecular energy calculations. After successful optimization the quantum-chemical molecular parameters are calculated.

The resulting prognoses are quantitative values of the target properties.

The top IDEF0 diagram is shown on figure 1. The box on the diagram represents the main technique's purpose, and arrows represent the data and objects used for this purpose achievement. The position of

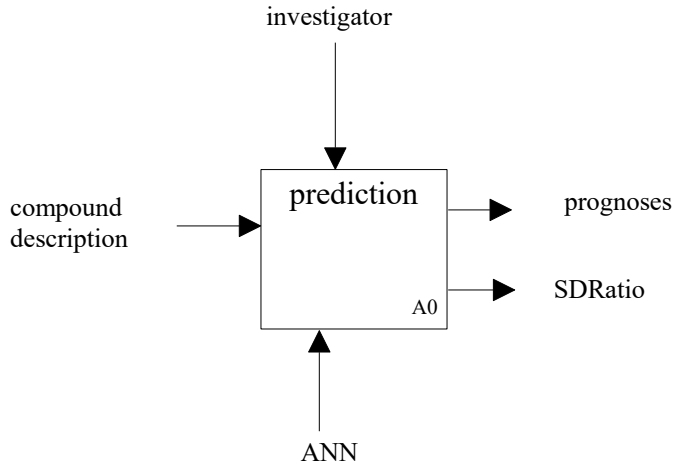


Figure 1. Top IDEF0 diagram

the arrow conveys its specific role. The controls enter the top of the box. The inputs enter the box from the left. The outputs of the operation leave the right-hand side of the box. Mechanism arrows join the bottom of the box.

2.2 The technique stages

The prediction technique consists of 5 stages. The stages are detailed at child IDEF0 diagram shown on figure 2. First, we have to divide the data set into the study set, the control set and the test set. The study set is used to built the ANN, the control set is used during training process and the test set is used to examine the predictive power of the ANN. It was found, that the power of the control and test sets should be less than 20% of the study set power. The study set and the control set are used at the next stage, the test set is used later during error analysis.

Second, for every study set we build the feed-forward ANN with 3 layers and 3 neurons in the hidden layer. We use STATISTICA Neural Networks [3] to implement the ANN training by one the most popular back-propagation Levenberg-Marquardt's algorithm.

Next stage is the examination of the ANN predictive power. The prognosis values for the test set are calculated by the ANN built at the previous stage. Then the differences between experimental and calculated values are analyzed. We use standard deviation coefficient (SDRatio) to estimate the error. This coefficient is calculated by STATISTICA Neural Networks as a ratio of the errors standard deviation to the data standard deviation. The small SDRatio (less then 0.1) means that the prognosis quality is appropriate. The SDRatio values are widely varied for different study sets, so we have to repeat the training procedure many times. Random or regular cross-validation methods are usually

applied to get the different sets. To estimate the predictive power of ANN for the compound set we calculate integrated error value.

One of the main problems of the ANN application for organic compound activity prediction is the dimension problem. Usually the dimension of the data set is many times less than the number of molecular descriptors. So ANN precisely describes the study set but not the other compounds. It was found that correlation between the descriptors usually is high (correlation coefficient is about 0.5-0.9). STATISTICA Neural Networks offers the genetic algorithm to reduce the number of the ANN inputs.

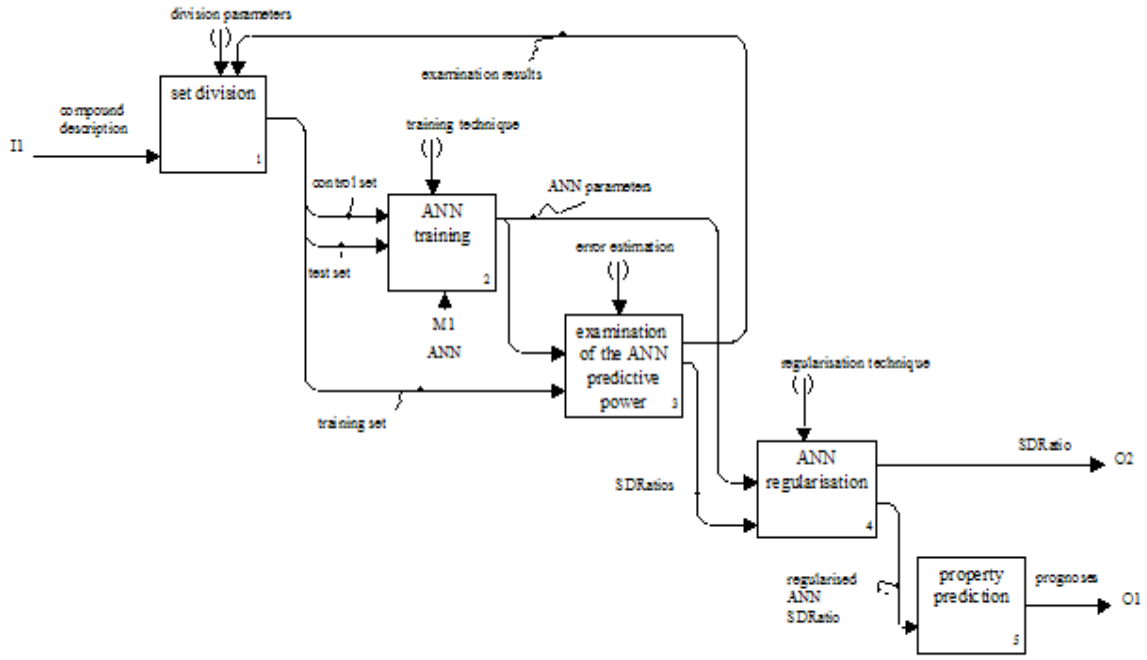


Figure 2. The prediction technique stages.

We provide the algorithm of more regularization of ANN set. The algorithm is detailed IDEF0 diagram at figure 3. We analyze the sets of input parameters for every ANN and choose the parameters that ate used in all of the networks or in the majority of them. Then we train ANN on the selected descriptors only and compare the errors. If the SDRatio of the regularized ANN is not more than STDRatios of the initial sets, the regularization is successful. Otherwise we have to change the quantity of the selected descriptors.

The final stage is the properties prediction. We use regularized ANN for this prediction. The

similar structure.

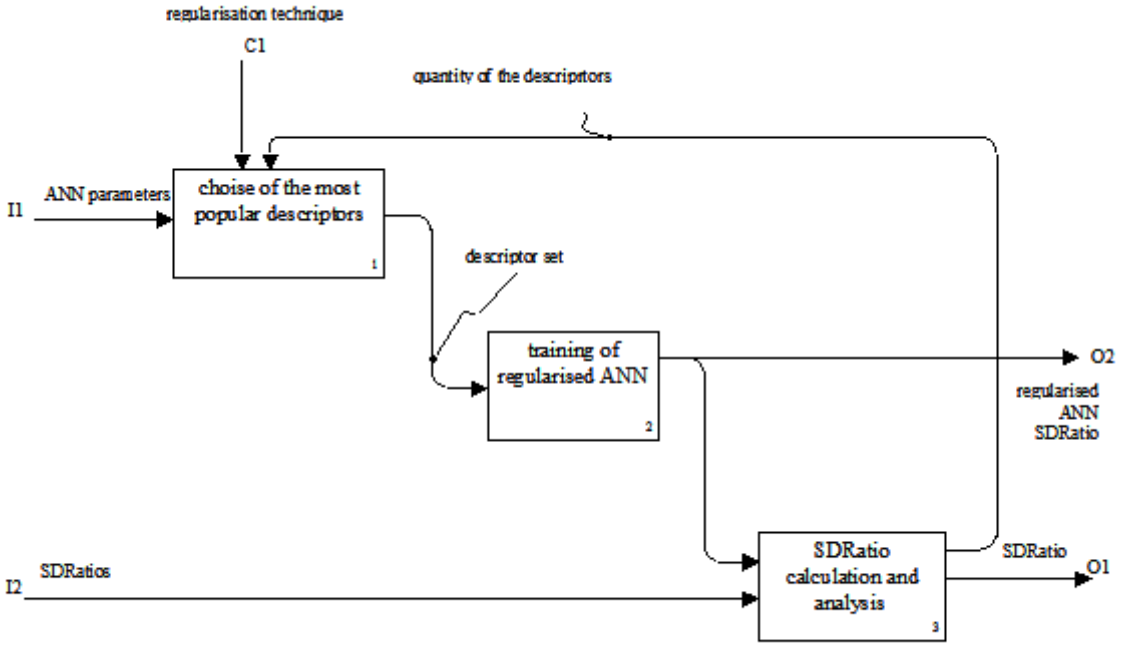


Figure 3. The regularization algorithm

We provide the algorithm of more regularization of ANN set. The algorithm is detailed IDEF0 diagram at figure 3. We analyze the sets of input parameters for every ANN and choose the parameters that ate used in all of the networks or in the majority of them. Then we train ANN on the selected descriptors only and compare the errors. If the SDRatio of the regularized ANN is not more than STDRatios of the initial sets, the regularization is successful. Otherwise we have to change the quantity of the selected descriptors.

The final stage is the properties prediction. We use regularized ANN for this prediction. The similar structure.

3. APPLICATIONS OF THE PREDICTION TECHNIQUE

The prediction technique was applied to 2 flexible organic compound series.

3.1 Ditiocarbamin acid derivatives

First set consists on 32 ditiocarbamin acid (DTCA) derivatives synthesized and tested on anti-tuberculosis activity in Russian State Antibiotic Center [4]. The feature of this set was 24 non-active compounds, so the traditional regression methods were useless for it. The molecule is flexible enough: the ‘nuclear’ part of the molecule has 4 bonds to be rotated. The ‘nuclear’ structure is shown at figure 4, and some compounds structures and activities are described at table 1.

Table 1. Compound structures and activities of some DTCA derivatives

R^1	R^2	R^3	R^4	R^5	R^6	R^7	Activity
NO_2	H	H	H	H	CH_2CH_3	CH_2CH_3	18
NO_2	H	NO_2	H	H	CH_2CH_3	CH_2CH_3	25
NO_2	H	F_3	H	H	CH_2CH_3	CH_2CH_3	19
NO_2	H	F_3	Cl	H	CH_3	CH_3	0
NO_2	H	F_3	H	NO_2	CH_3	CH_3	0

First, the exhaustive search with 1 control compound and 1 test compound. Mean SDRatio

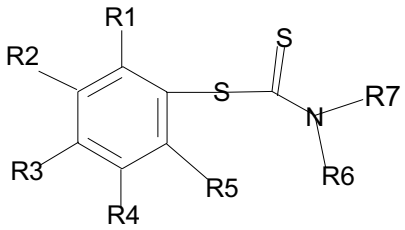


Figure 4. The structure of DTCA derivatives

value of all of these ANN was 0.1. We regularize the sets of ANN with the same control

compound. The results for first 10 compounds are presented at table 2. Zero value of SDRatio in the table means that the value is less than 0.0001. The mean SDRatio value was 0.0006. So the regularization improves the models quality.

Later, the random cross-validation technique was applied to these compounds with 1 control compound and 3 test compounds. Some results are presented at table 3. The study set was reduced but the mean SDRatio value of 30 experiments was 0.04.

The activity prognoses of the II, XXVIII, XXX, XXXII compounds were the best. These compounds are to be tested as the base structures for the anti-tuberculosis drugs.

3.2 Tetrahydroimidazobenzodiazipenone derivatives

Tetrahydroimidazobenzodiazipenone derivatives (TIBO) are known as the non-nucleoside reverse transcriptase inhibitors that can be used in anti-AIDS drugs. The data set consists on 49 compounds taken from [5,6,7,8]. The 50% inhibitory concentration of the TIBO was an activity under study. The structures and activities of the compounds were found form literature. Figure 5 shows the ‘nuclear’ structure (the cycle of 7 atoms is not planar). Table 4 describes some compounds.

Table 2. Parameters of some regularized ANN

Test compound	Quantity of input variables	SDRatio
I	23	0,0031
II	21	0
III	22	0,0012
IV	24	0
V	23	0,0006
VI	21	0,0005
VII	22	0,0010
VIII	23	0,0002
IX	24	0
X	23	0

Table 3. Results of the random cross-validation of the DTCA compounds

Experiment	Control set	Test compound	SDRatio
1.	XI X XX	II	0
2.	XI XXII XXIV	XXIII	0,02
3.	XVIII XXVII XXV	I	0
4.	X XXI XXIV	XV	0
5.	XI XXII XXIV	XV	0
6.	XX XXIV XXV	III	0
7.	VIII XVIII XXVIII	XXIII	0,16
8.	IV VIII XXI	III	0,36
9.	XIV XXIV XXVIII	IV	0
10.	XVII XXV XXVIII	VIII	0

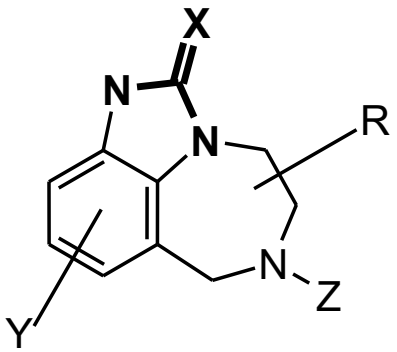


Figure 5. The structure of TIBO derivatives

Table 4. Compound structures and activities of some TIBO derivatives

R	X	Y	Z	$IC_{50}(\mu M)$	$-\log(IC_{50}(\mu M))$
H	S	8-Cl	dimethylallyl	0,046	1,337
H	S	9-Cl	dimethylallyl	0,16	0,796
7-Me	O		dimethylallyl	12,1	-1,083
7-Me	O	8-Cl	dimethylallyl	0,145	0,839
7-Me	O	9-Cl	dimethylallyl	0,16	0,796

We use random cross-validation technique with 2 control and 2 test compounds. Some results are presented at table 5. The error of regularized ANN for every set is less than the error of unregularized ANN, and the SDRatio value after the regularization shows that the prediction power is appropriate.

Table 5. Results of the random cross-validation of the TIBO compounds

Control set	Test set	SDRatio (unregulized ANN)	SDRatio (regulized ANN)
h-III, s-XVI	h-V, i-IX	0,1	0,02
s-XI, h-IV	t-I, i-I	0,3	0,09
s-XV, p-IV	s-X, p-VII	0,1	0,06
s-I, i-VIII	s-XV, h-VII	0,4	0,31
i-VII, h-VII	s-II, s-XX	0,2	0,03
i-XI, s-XVI	s-III, h-III	0	0
s-II, s-XI	p-IV, h-III	0,2	0,07
s-XX, i-I	s-V, s-XI	0,2	0
s-I, t-II	s-VIII, p-VI	0,2	0,14
s-IV, i-VII	i-X, h-II	0,1	0,02
Mean value		0,18	0,075

4. CONCLUSION

The technique of flexible organic compounds properties prediction based on ANN was been provided and described by IDEF0 diagrams. The algorithm for the ANN set regularization was developed. The results of the tests on 2 flexible organic compound series show the appropriate predictive power of the regularized ANN models. Further investigations are related with the 3D molecular structure selection.

REFERENCES

1. Zupan J., Gasteiger J. Neural Networks in Chemistry and Drug Design. Second Edition. Wiley-VCH, Weinheim 1999: 380.

2. t1.chem.umn.edu/mopac/

3. www.statsoft.com

4. Граник В.Г, Макаров В.А. Синтез и свойства производных тиокарбаминовых кислот. Успехи химии 1998; 11: 1013-1032.

5. Silverman B.D., Platt D.E. Comparative Molecular Moment Analysis: 3D-QSAR without Molecular Superposition. Journal of Medicinal Chemistry 1996; 39-11: 2129-2140.

6. Kukla M.J., Breslin H.J., Diamond C.J. et al. Journal of Medicinal Chemistry 1991; 34: 3187.

7. Hannongbua S., Pungpo P., Limtrakul J., Wolschann P. Quantitative structure-activity relationships and comparative molecular field analysis of TIBO derivatised HIV-1 reverse transcriptase inhibitors. Journal of Computer-Aided Molecular Design 1999; 13: 563-577

8. Eriksson Mats A. L., Pitera J., Kollman P. Prediction of the Binding Free Energies of New TIBO-like HIV-1 Reverse Transcriptase Inhibitors Using a Combination of PROFEC, PB/SA, CMC/MD and Free Energy Calculations. Journal of Medicinal Chemistry 1999; 42-5: 868-881.