



Quantitative structure activity relationship indanylacetic acid derivatives models based on a novel machine learning method

Peijian Zhang^a, Hongzong Si^b, Yun Bo Duan^b, Gengxin Sun^a, Kejun Zhang^c and Hongling Zhai^d

^aInternational College, Qingdao University, Qingdao, Shandong, China

^bInstitute for Computational Science and Engineering, Laboratory of New Fibrous Materials and Modern Textile, the Growing Base for State Key Laboratory, Qingdao University, Qingdao, Shandong, China

^cDepartment of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, China

^dDepartment of Chemistry, Lanzhou University, Lanzhou, Gansu, China

ABSTRACT

Quantitative structure activity relationship (QSAR) models for prediction of the EC_{50} of the peroxisome proliferator-activated receptors (PPARs) have been developed basis on the linear heuristic method(HM). Molecular descriptors were used to represent the characteristics of compounds. HM was used to pre-select the whole descriptor sets and to build the linear model. The new compounds were designed according to the QSAR models. The same descriptors were applied in the model and the satisfied EC_{50} values were obtained for the designed compounds. The selected descriptors will help for new drug design and the models are available for predicting the EC_{50} of the new drugs.

Key words: QSAR; Heuristic Method (HM); Peroxisome Proliferator-Activated Receptors (PPARs)

INTRODUCTION

In developed countries, chronic diseases such as diabetes, obesity, atherosclerosis and cancer are the most frequent reasons which cause of death. The peroxisome proliferator-activated receptors (PPARs) are a group of nuclear receptors (NRs) that control many cellular metabolic processes.

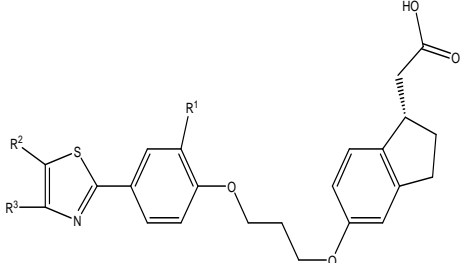
Recently, a new group of PPARs with thiazolyl-phenoxy were synthesized [1]. The EC_{50} (a quantal dose response curve represents the concentration of a compound) of all compounds were evaluated for human PPAR α and human PPAR δ in vitro potency by fluorescence resonance energy transfer (FRET) assays. PPAR γ activity was assessed using a cell-based functional assay (IRBA) in mouse T3-L1 cells [1, 2]. However, these methods are laborious, expensive, or time-consuming and require a sufficient quantity of the pure compounds. Therefore, there need suitable way to high-throughput screening of millions of compounds. QSAR is a potential and useful technique to estimate the EC_{50} , especially for the compounds that are not easy to test.

QSAR study is a key step for new drug design and screen. The HM is a novel machine learning method [3], which has been successfully used to predict the evaporation estimation [4] and the cement strength [5]. In the present work, the HM was utilized to set up the QSAR model for a new class PPARs compounds. Basis on the model, we designed a group structures and predicted the EC_{50} by the model. Therefore, three compounds have been found the good EC_{50} values.

EXPERIMENTAL SECTION

2.1. Data preparation

The experimental values for the EC₅₀ of PPARs were taken from the literature [1]. The data set was randomly separated into a training set of 28 compounds and a test set of 12 compounds in PPAR α and PPAR δ groups, and a training set of 27 compounds and a test set of 12 compounds PPAR γ group. The training set was used to build the model and the test set was employed to evaluate the prediction ability of the model (Table 1).

Table 1. Experimental and predicted log(EC₅₀) of 4-Thiazolylphenyl Analogs by HM


No.	R ¹	R ²	R ³	PPAR α log(EC ₅₀)	
				Exp.	HM
1	n-Pr	H	H	2.05	2.17
2	OMe	H	H	2.41	3.30
3	OMe	Me	H	2.43	2.96
4	H	H	H	3.11	2.66
5	n-Pr	H	Et	2.17	1.85
6	OMe	H	H	2.40	2.31
7	n-Pr	H	t-Bu	1.94	1.96
8	n-Pr	H	CF ₃	2.29	2.27
9	OMe	H	H	2.64	2.69
10	H	H	H	3.10	2.62
11	OMe	Me	Me	1.92	2.26
12	H	H	H	2.05	2.51
13	n-Pr	H	H	2.52	1.88
14	OMe	H	H	1.72	2.28
15	H	H	H	2.59	2.34
16	n-Pr	H	H	1.63	1.73
17	OMe	H	H	1.64	2.11
18	n-Pr	H	H	2.00	2.08
19	OMe	H	H	2.52	2.33
20	OMe	H	H	2.72	2.44
21	H	H	H	3.40	3.13
22	n-Pr	COCH ₃	Me	2.24	2.42
23	OMe	H	H	2.75	2.82
24	n-Pr	CONMe ₂	Me	2.15	2.37
25	OMe	H	H	2.90	2.93
26	H	H	H	4.00	4.19
27*	n-Pr	COOH	Me	3.23	3.07
28*	OMe	H	H	4.00	3.42
29*	n-Pr	COOH	CH ₂ OH	3.81	3.82
30*	n-Pr	H	CH ₂ CO ₂ H	3.39	3.28
31*	H	H	OMe	3.76	3.09
32*	OMe	H	H	2.88	2.87
33*	H	H	H	2.00	2.70
34*	n-Pr	H	OEt	2.02	2.08
35*	OMe	H	H	3.13	2.61
36*	n-Pr	H	H	1.67	1.87
37*	OMe	H	Oi-Pr	2.26	2.31
38*	n-Pr	Me	OEt	1.79	2.05
39	OMe	Me	OEt	2.18	2.11
40	OMe	Et	OEt	2.72	2.25

The star "*" represents the test set.

2.2. Calculation of the descriptors

To obtain a QSAR model, compounds are often represented by the molecular descriptors. All molecules were drawn into Hyperchem [6] and pre-optimized using MM+ molecular mechanics force field. A more precise optimization had been done with semi-empirical AM1 method in MOPAC [7]. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient is 0.01. The MOPAC output files were used by the CODESSA program [8, 9] to calculate five classes of descriptors: constitutional, topological, geometrical, electrostatic, quantum chemical. CODESSA combines diverse methods for quantifying the structural information about the molecule with advanced statistical analysis to establish molecular structure–property/activity relationships. CODESSA had been applied successfully in a variety of QSAR analyses [10-13].

2.3. Development of linear model by the HM [10-13]

Once the molecular descriptors are generated, the HM in CODESSA is used to pre-select the descriptors and build the linear model. The advantages of the HM are the high speed and no software restrictions on the size of the data set. The HM can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. The details of selecting descriptors are as follows: First of all, all descriptors are checked to ensure that values of each descriptor are available for each structure. Descriptors for which values are not available for every structure in the data are discarded. Descriptors having a constant value for all structures in the data set are also discarded. Thereafter all possible one-parameter regression models are tested and the insignificant descriptors are removed. As a next step, the program calculates the pair correlation matrix of descriptors and further reduces the descriptor pool by eliminating highly correlated descriptors. The details of validating intercorrelation are (a) all quasi-orthogonal pairs of structural descriptors are selected from the initial set. Two descriptors are considered orthogonal if their inter-correlation coefficient r_{ij} is lower than 0.1; (b) CODESSA uses the pairs of orthogonal descriptors to compute the bi-parametric regression equations; (c) to an MLR model containing n descriptors, a new descriptor is added to generate a model with $n+1$ descriptors if the new descriptor is not significantly correlated with the previous n descriptors; step (c) is repeated until MLR models with a prescribed number of descriptors are obtained. The goodness of the correlation is tested by the square of coefficient regression (R^2), square of cross-validate coefficient regression (R_{CV}^2), the F-test (F), and the standard deviation (s^2).

RESULTS AND DISCUSSION

The HM was used to develop the linear model for prediction the EC_{50} of PPARs basis on the calculated structural descriptors. The correlation coefficient value of each the two descriptors are lower than 0.80, which means that the descriptors are independent in the analysis. The correlation model was given as follows:

$$EC_{50(PPAR\alpha)} = -5.59 - 54.21 FNSA + 416.38 PMI + 1.03 FPSA$$

$n = 28$, $R^2 = 0.69$, $F = 27.33$, $RMS = 0.14$, where $FNSA$, PMI and $FPSA$ represent FNSA-3 Fractional PNSA (PNSA-3/TMSA, Quantum-Chemical PC), principal moment of inertia A and FPSA-2 Fractional PPSA (PPSA-2/TMSA, Quantum-Chemical PC), respectively.

CONCLUSION

In this work, we applied linear and non-linear models for the prediction of EC_{50} value of a set of 40 PPARs. The proposed linear model could give the satisfied QSAR models for three groups' compounds. The results of this work indicate that the HM is a very promising tool. These models will assist the future drug design for PPAR receptors and the EC_{50} can be predicted by the corresponding model.

REFERENCES

- [1] Rudolph, L.B. Chen, D. Majumdar, W.H. Bullock, M. Burns, T. Claus, F.E. Cruz, M. Daly, F.J. Ehrgott, J.S. Johnson, J.N. Livingston, R.W. Schoenleber, J. Shapiro, L. Yang, M. Ma, X. Tsutsumi, *J Med Chem.* 50 (2007) 984-1000.
- [2] Shujun Kong, Jianqing Hou, Min Xia, Ying Yang, Anli Xu, Qing Tang, *Cancer Cell Research* 2014 1(2) 37-41.
- [3] C. Ferreira, *Gene Expression Programming in Problem Solving. Soft Computing and Industry-Recent Applications. Applications, Springer-Verlag*, 2002, pp, 635-654.
- [4] T. Ozlem, M.E. Keskin, *J Appl Sci.* 5 (2005) 508-512.
- [5] A. Baykasoglu, T. Dereli S. Tanis, *Cement and Concrete Research.* 34 (2004) 2083-2090.

- [6] HyperChem 4.0, Hypercube, Inc., **1994**. 26, pp, 5-14.
- [7] J.P.P. Stewart, MOPAC 6.0, *Quantum Chemistry Program Exchange, QCPE*, No. 455, Indiana University, Bloomington, IN **1989**.
- [8] A.R. Katritzky, V.S. Lobanov M. Karelson, *Reference Manual, version 2, University of Florida: Gainesville, FL*, **1994**.
- [9] A.R. Katritzky, V.S. Lobanov M. Karelson, *Chem. Soc. Rev.* 24 (**1995**) 279-287.
- [10] Lei Liu, Xiurui Han, Hongzong Si and Lianhua Cui, *J. Comput. Sci. Eng.* 12 (**2014**) 438-444.
- [11] H.Z. Si, T. Wang, K.J. Zhang, Z.D. Hu, B.T. Fan, *Bioorgan Med Chem.* 14 (**2006**) 4834-4841.
- [12] Jiazhong Li, Juanjuan He, Beilei Lei, Huanxiang Liu and Paola Gramatica, *J. Comput. Sci. Eng.* 9 (**2013**) 326-355.
- [13] A.R. Katritzky, R. Petrukhin, R. Jain, M. Karelson, *J Chem Inf Comput Sci.* 41 (**2001**) 1521-1530.