Journal of Computational Science &Engineering 41 (2019) 1056-1061

Available at http://www.asocse.org ISSN 1710-4068

QSAR Study for Potential Inhibition of Aromatic Ring

Ke Wang[1], Hongzong Si[2,*]

[1]Department of Data Science and Software Engineering,Qingdao University Medical College,266071,China;

[2]Institute for Computational Science Engineering, Qingdao University, Qingdao, 266071, China

Abstract: **With the development of agricultural and chemical research, it has been found that some of derivatives are usually with higher plant growth inhibition than those natural. The purpose of our research is to acquire relatively effective structural determinants to predict the inhibition of the new aromatic ring bearing many different kinds of substituents by establishing Quantitative structure activity relationship (QSAR). In this work, we used CODESSA software to calculate the molecular descriptors to build mathematical relationship between physicochemical properties and standardize experimental data. We accessed the gene expression programming by APS software to predict the inhibition of the derivatives. As a result, the accuracy of gene expression programming in training and test sets are respectively 97.778% and 86.667%. Thus, the exactitude manifest the QSAR we obtained is a effective model.**

Keywords: QSAR; Aromatic ring; Eene expression programming

Received 5 May 2019, Revised 7 June 2019, Accepted 10 June 2019

*Corresponding Author: Hongzong Si, sihz03@126.com

1. Introduction

Aromatic rings (also known as aromatic compounds or arenes) are hydrocarbons which contain benzene, or some other related ring structure. It has been reported that aromatic ring lignano-9,9'-lactonea is a structure of Matairesinol[1], which is well-known dietary lignan biosynthesized in vegetables, which was reported to be related to the cytotoxicity of arctigenin[2], human monoamine[3] and the GABA transporter activities, antiinflflammatory[4] activities of hinokinin, the suppression of neuroinflflammation, and the IgE-suppressive activity of matairesinol[5]. It shows the higher cytotoxicity than arctigenin. Because of the recent theory that most of the synthetic growth inhibitors usually have higher inhibitory activity, this kind of specific molecule inhibitor aroused my impulse to research into more details.

As the increasing methods are built to design the structure of drugs, some of the these have gradually exhibited weaknesses. Assaying the large number of derivatives by experiment in laboratory is a heavy losses not only for the experimental natural resources but also the time we consuming. Obviously, executing such test for hundreds of new chemical compounds is a laborious, costly and even unachievable method. Quantitative structure activity relationship (QSAR) is a classification models relating the predictor variables, like physicochemical properties, to a categorical value of the response variable. In the past decades, establishing a bionic mathematics calculation model has solved many prediction problems successfully. Meanwhile, it can reduce the potentially dispensable plants testing requirements and improving the efficiency. In order to advance to explore more sophisticated models, we use several computational and statistical software and tools. Gene expression programming (GEP) was an automatic programming approach introduced by Ferreira. These computational programs are complex tree structures learning and adapting by changing their sizes, shapes, and composition. It overcomes certain limitations of genetic algorithms and genetic programming by working with two elements, the chromosome and the expression tree. Each physical and chemical parameter of aromatic ring is as a gene unit in the gene expression programming[6][7]. After complex calculation, they are all weaved to a multivariate nonlinear equation. All the advantages of GEP make it a incomparably superior method for solving classification problems and have been applied to chemistry analysis for years.

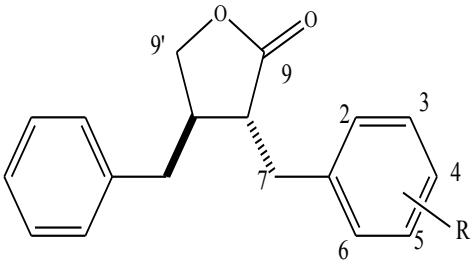
2. Methodology

2.1 Data Preparation

31 structures and IC_{50} values for aromatic ring lignano-9,9'-lactonea were obtained from a literature and listed in 2 groups (Table1 and Table 2). Table 1 is based on the experiment towards Lettuce Seedling Shoots and Roots and Table 2 is the result of the inhibition for Italian Ryegrass Seedling Shoots and Roots. They were cataloged by the value of IC_{50}. This quantitative measure indicates how much of a particular drug is needed to inhibit a given biological process by half. The smaller the IC_{50} is, the more potent the drug is. So we took 10 compounds and 13 compounds separately from two group whose value of IC_{50} shows that they can be significantly against the roots, While those whose value of the IC_{50} is over 1.0 were eliminated. Finally, we set 1 for inhibitor, 0 for non-inhibitor. At last, 62 compounds were divided into 46 training sets and 16 corresponding test sets.

2019@ The American Computational Science Society. All rights reserved.

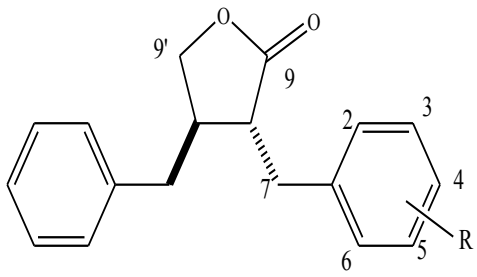
Table 1. Plant Growth Inhibitory Activities of 7−37 against Lettuce Seedling Shoots and Roots[a]



Compound	R	Inhibition(exp)	Inhibition(GEP)
1	H	0	0
2	2-OH	1	1
3	2-OCH_3	0	0
4	2-CH_3	0	0
5	2-F	0	0
6	3-OH	1	1
7	3-OCH_3	0	0
8	3-CH_3	0	0
9	3-F	0	0
10	4-OH	1	1
11	4-OCH_3	0	0
12	4-CH_3	0	0
13	4-F	0	0
14	3,4-OCH_2O	0	0
15	3-OCH_3, 4-OCH_3	0	0
16	3-OCH_3, 4-OH	0	0
17	3-$O(CH_2)_3CH_3$, 4-OH	0	0
18	2-OH, 3-OH	1	1
19	2-OH, 4-OH	1	1
20	2-OH, 5-OH	1	1
21	2-OH, 6-OH	1	1
22	3-OH, 4-OH	1	1
23	2-OH, 5-OCH_3	0	0
24	2-OH, 3-CH_3	0	0
25	2-OH, 5-CH_3	0	0
26	2-OH, 3-F	0	0
27	2-OH, 4-F	1	1
28	2-OH, 5-F	1	1
29	2-CH_3, 4-F	0	0
30	2-CH_3, 5-F	0	0
31	2-F, 5-F	0	0

2019@ The American Computational Science Society. All rights reserved.

Table 2. Plant Growth Inhibitory Activity of 7−37 against Italian Ryegrass Seedling Shoots and Roots[a]



Compound	R	Inhibition(exp)	Inhibition(GEP)
1	H	0	0
2	2-OH	1	1
3	2-OCH_3	0	0
4	2-CH_3	0	0
5	2-F	0	0
6	3-OH	1	1
7	3-OCH_3	0	0
8	3-CH_3	0	0
9	3-F	0	0
10	4-OH	1	1
11	4-OCH_3	0	0
12	4-CH_3	0	0
13	4-F	0	0
14	3,4-OCH_2O	0	0
15	3-OCH_3, 4-OCH_3	0	0
16	3-OCH_3, 4-OH	0	0
17	3-$O(CH_2)_3CH_3$, 4-OH	0	0
18	2-OH, 3-OH	1	1
19	2-OH, 4-OH	1	1
20	2-OH, 5-OH	1	1
21	2-OH, 6-OH	1	1
22	3-OH, 4-OH	1	1
23	2-OH, 5-OCH_3	1	1
24	2-OH, 3-CH_3	0	0
25	2-OH, 5-CH_3	1	1
26	2-OH, 3-F	1	1
27	2-OH, 4-F	1	1
28	2-OH, 5-F	1	1
29	2-CH_3, 4-F	0	0
30	2-CH_3, 5-F	0	0
31	2-F, 5-F	0	0

2019@ The American Computational Science Society. All rights reserved.

2.2 Calculation of Molecular Descriptors

In order to use molecular descriptors to build the relationship between the biological activity and molecular structure parameters, 30 compounds were drawn using ChemDraw. And we used Hyperchem 7.5 software to optimize the derivatives' structure.The calculation used MM+ molecular mechanics force field to pre-optimize, and the semi-empirical AM1 method[8] to get more precise geometrical optimization about the lowest energy conformation. Semi-empirical quantum chemistry methods, based on the Hartree–Fock formalism, can make many approximations and acquire some parameters from empirical data. They are very suitable for computational chemistry and analyzing properties relating to physical chemistry of large molecules, since without the accurate estimations, the full Hartree–Fock method[9] is too expensive. In this way, semi-empirical AM1 calculation can be successfully taken for the study of QSAR. The molecular structures would be unceasingly optimized by Polak-Ribiere algorithm until the root mean square gradient reached to 0.01[10]. Next, MNO files containing kinds of information of molecular system were generated by MOPAC calculation. The built model performance and results accuracy are strongly depended on the way that descriptor was performed. The CODESSA program, developed by Katritzky group, calculates a large number of quantitative descriptors based on the molecular structural information[11] and the chemical information is transformed into mathematical form after codification . It can give the following kinds of descriptors: constitutional descriptors, topological descriptors, geometrical descriptors, electrostatic descriptors, and quantum-chemical descriptors. After being calculated in APS software, they finally generated 420 descriptors according to the characteristics.

2.3 The selection of the descriptor and Theory of GEP

Nevertheless, there are an extensive matrix of molecular descriptors, such as molecular weights, rotatable bonds, the number of hydrogen bond donors and acceptors, or some pre-experimental or pre-calculated properties. Some descriptors generated for each compound and encoded similar information about the molecule of interest. Therefore, it was requisite to test each descriptor and eliminated those that show high correlation. After calculation and correlation analysis of the values of the R^2 and F statistic, five descriptors were chosen.

Gene expression programming (GEP)[12], like genetic algorithms (GAs) and genetic programming (GP)[13], is a evolutionary and genetic algorithm. It bases on the structure and function of biological gene. It uses populations of individuals, selects them according to fitness, and introduces genetic variation using one or more genetic operators . The elementary differences between three algorithms are in the properties of the individuals: in GAs the individuals are linear with fixed length[14]; in GP the individuals are nonlinear entities and have different sizes and shapes[15]; in GEP the individuals are encoded as linear strings of fixed length which are later expressed as nonlinear entities of different sizes and shapes. The search area of GEP is divided from the solution space, which leads to lots of advantages, such as the independent search of the genome space. Gene expression programming uses the chromosomes composed of genes structurally organized in a head and a tail. The head is used to encode functions for the expression. And the tail is a container of extra terminal symbols. It can be used if there aren't enough terminals in the head to provide arguments for the functions. Thus, the head can contain functions, variables and constants, but the tail can contain only variables and constants. The number of symbols in the head of a gene is specified as a parameter for the analysis. The number of symbols in the tail is determined by the equation

$t = h*(MaxArg-1)+1$

Where "t" is the number of symbols in the tail, "h" is the number of symbols in the head, and MaxArg is the maximum number of parameters required by any function that is allowed to be used in the expression. The key to GEP's ability to vary quickly is the way it encodes symbols in genes. The chromosomes function as a genome, after being modified by various means of mutation, transposition, root transposition, gene transposition, gene recombination, they will be encoded as expression trees. For example, the following simple mathematical expression:

$$\sqrt{(a+b)\times(c-d)}$$

It can be transformed into an expression tree(ET) shows as Figure 1.

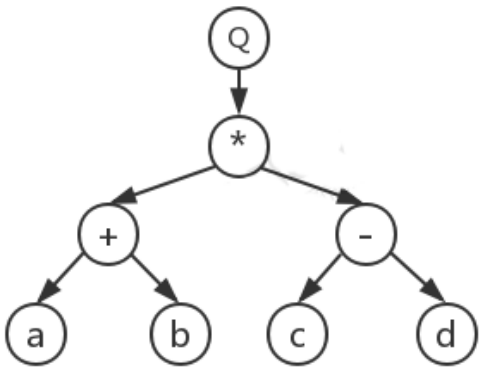


Figure 1. Schematic diagram of ET

Moreover, parameters in the operation relationship were continuously used to construct various models until get the best results. In GEP, fitness is based on how well an individual models the data. For classification problems with a categorical established variable, fitness can be evaluated by the number of

2019@ The American Computational Science Society. All rights reserved.

correct predictions. Evolution stops when the fitness of the best individual in the population reaches some limits which are specified for the analysis, or when a vested generations have been created or a maximum execution time limit is reached. This complex functions can improve the prediction ability of QSAR model. All the fitness functions produce fitness scores from 0.0 to 1.0 with 1.0 being ideal fitness.Roulette-wheel sampling is used to find the optimal solution when individuals moving on to the next generation. The schematic diagram displays as Figure.2.

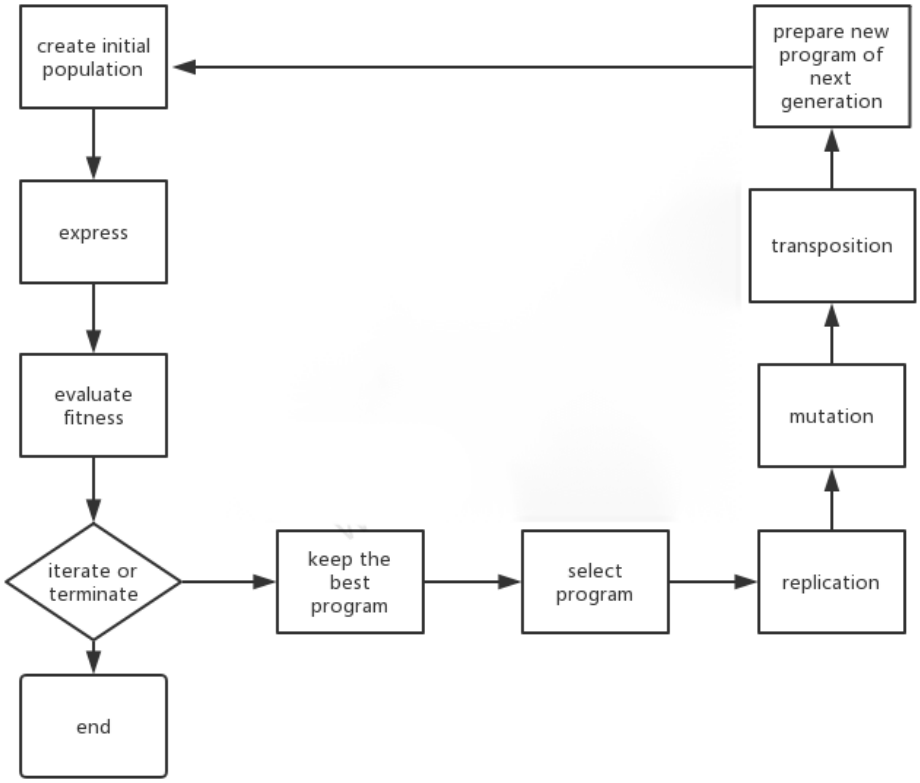


Figure 2. The flowchart of gene expression algorithms

3. Results and Discussion

62 organics includes 23 inhibitive and 39 non-inhibitive compounds. The number of inhibitors and non-inhibitors is 46 in training set and 16 in test set. The setting of the function is : {+,-,*,/}

5 Groups descriptors were used to build GEP model in the Automatic Problem Solver 3.0 (Gepsoft Limited company, Bristol, United Kingdom). After converting the Java function, equation shows in Figure 3.

$$F_{(x)} = 3x_1 + 3x_2 - 2x_3 + 2x_4 - \frac{x_2}{x_2 - x_3} + (x_1^{\,2} + x_0) \cdot x_3 x_4$$

Figure 3. Equation 1

The variables x_0, x_1, x_2, x_3, x_4 represents PPSA-3 Atomic charge weighted PPSA [Quantum - Chemical PC],Polarity parameter/square distance, RNCG Relative negative charge (QMNEG/QTMINUS) [Quantum-Chemical F], Tot dipole of the molecule, Max partial charge (Qmax).

The mechanistic interpretation of the endpoint of interest may not always provide easily interpretable information. Finally, we choose PPSA-3(the atomic charge weighted partial positively charged surface area[16]) Atomic charge weighted PPSA [Quantum - Chemical PC], Polarity parameter/square distance, RNCG Relative negative charge (QMNEG/QTMINUS) [Quantum - Chemical F], Tot dipole of the molecule and Max partial charge (Qmax) to represent the rational explanations of their effects on inhibition. The next step, after precise analysis, the QSAR model to predict the inhibition of the synthetic chemicals is principally founded. Subsequently, we tested and evaluated the model with different model parameters which provide the potent meaning under the relationship of different key molecular descriptors attached to the QSAR model.

4. Conclusions

Building QSAR models for aromatic chemicals were essential in the study of plant growth inhibition. QSAR

2019@ The American Computational Science Society. All rights reserved.

provides an innovative, eco-friendly and reliable method to explain the potential biological properties and predict the inhibition. In this study, the unique advantage of GEP displayed remarkably that it can establish accurate and easily-interpreted mathematical model. Based on the model, more new compounds with higher potential inhibition can be designed in the future. Admittedly, the model still need to be further examined and tested. Therefore, all the approaches and methods in this study will be useful guidelines for other research.

References

[1] J.L. Penalvo, H. Adlercreutz, M. Uehara, A. Ristimaki, S. Watanabe. J. Agric. Food Chem, 56(2008) 401-409.

[2] K. Umehara, A. Sugawa, M. Kuroyanagi, A. Ueno, T. Taki. Chem. Pharm. Bull, 41(1993) 1774−1779.

[3] J. M.V. Timple, L.G. Magalhaes, O.V. Mortensen, A.C.K. Fontana. J.Nat.Prod, 76(2013) 1889-1895.

[4] T.C. Lima, R. Lucarini, A.C. Volpe, C.Q.J. de Andrade, A.M.P. Souza, P.M. Pauletti, A.H. Januaŕio, G.V. Símaro, J.K. Bastos, W.R. Cunha, A. Borges, R.da S. Laurentiz, V.A. Conforti, R.L.T. Parreira, C. H.G. Borges, G.F. Caramori, K. F. Andriani, M. L. A. e Silva. Bioorg. Med. Chem, 27(2017) 176-179.

[5] Y. Satoshi, Y. Yasuka, N. Asuka, N. Hisashi. Journal of Agricultural and Food Chemistry, 66(2018) 4551-4558.

[6] J. Joanna, J. Piotr. Expert Systems with Applications, 38(2011) 10932-10939.

[7] A.E. Seyed, E-N. Hamid. Construction and Building Materials, 218(2019) 8-27.

[8] F.B.G. Paulo, R.L. Paolo. Chemical Physics Letters, 304(1999) 438-444.

[9] R. Rencsok, K.A. Jackson, T.A. Kaplan, J.F. Harrison, M.R. Pederson. Chemical Physics Letters, 262(1996) 207-212.

[10] W.J. Robert, A.P. Rex. Journal of Molecular Structure, 39(1995) 1174-1179.

[11] R.C. Fernandes. Moniac, Huang Xiaoming, C.L. Abbenhuis. Hendrikus, J.M. Hensen. Emiel. International Journal of Biological Macromolecules, 123(2019) 1044-1051.

[12] Z. Jing, Y. Gui, T. Junlong, W. Zhi, M. Dongwei. Applied Surface Science, 475(2019) 863-872.

[13] T. Kailin, L. Tonghua. Chemometrics and Intelligent Laboratory Systems, 64(2002) 55-64.

[14] S. Ali, H. Mohsom, S. Majid. Optics & Laser Technology, 113(2019) 104-115.

[15] K. Murat. Computers & Geosciences, 37(2011) 1318-1323.

[16] A.R. Katritzky, V.S. Lobanov, M. Karelson, Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual, Version 2.0, 1994.

2019@ The American Computational Science Society. All rights reserved.