Journal of Computational Science & Engineering 21 (2015) 658-661

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

QSAR Study on the Safety Evaluation of Compounds

Fucheng Song, Lianhua Cui, Jun Liu, Jiahong Zheng, Fenglin Sun, Yaqiong Luo, Qing Liu, Bing Ge, Hongzong Si*

Department of Public Health, the Medical College of Qingdao University, Qingdao, Shandong 266071, China

Abstract: **The quantitative structure-activity relationship (QSAR) is using the theoretical calculation and statistical analysis software tools to study the chemical structure and biological effects of quantitative relationship. It is based on a traditional structure-activity relationship and the combination of empirical equation mathematical methods commonly used in physical chemistry. Gene expression programming (GEP) is a relatively new evolutionary computation algorithm, which is a kind of good data mining and modeling tools. Heuristic method (HM) has strong ability of function mining and can rapidly find the answer in complex problem. QSAR model can be used to predict the activity of untested compounds or the physical and chemical properties. On the other hand, QSAR can reveal the structural factors which affect the properties of security risk of compounds, and further better understand the mechanism of action of the toxicity.**

Keywords: QSAR; Gene expression programming; Heuristic method; Safety evaluation

Received 15 October 2015, Revised 23 November 2015, Accepted 25 November 2015

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

1. Introduction

The quantitative structure-activity relationship (QSAR) is the use of theoretical calculation and statistical analysis software tools to study the chemical structure and biological effects of quantitative relationship that is a set of biological, chemical and statistics for the integration of a comprehensive technology. The theoretical basis of QSAR is the ability to influence the chemical reaction rate will be embodied in different response of active or quantifiable index. The structure-activity relationship begin from a homologue of the similar elements composition, spatial structure and biological activity, find out the homologue quality, and the relationship between chemical structure and toxicity in the composition of the most closely. So in prediction of the effects of health hazards of exogenous compounds can begin from the chemical structure of harmful substances characteristics to assess the potential toxicity of compound [1-4]. It is well known that a development of new drugs cost enormous efforts, material and financial resources. But with the rapid development of computer technologies, the increasing number of methods is utilized to predict the activities of the chemicals using their structural properties and screen a large number of alternative compounds [5-7]. The QSAR method is a study of physicochemical properties of compounds based on the molecular descriptors which are calculated by their structures. QSAR has become an efficient way for drug discovery. QSAR has been proven successfully in the discovery of lead compounds [8,9].

QSAR is based on a traditional structure-activity relationship and the combination of empirical equation mathematical methods commonly used in physical chemistry. Its theory history can be traced back to Crum-Brown equation in 1868. The compounds of physiological activity could use a function of the chemical structure, but not clear function model was set up. The earliest equation could implement the quantitative structure-activity relationship that was Hansch equation put forward by Hansch in 1962. Subsequently, Hansch and Japan visiting scholar fujita mori improved Hansch equation and introduced the indicator variables, parabolic model and bilinear model correction that making equation prediction ability improved. Almost at the same time of the Hansch method published, Free and others published Free-Wilson method. This method directly used the molecular structure as variables regression analysis to carry out on the physiological activity. Hansch and Free - Wilson methods both take the molecules as a whole to consider the nature, could not detailed the relationship between the three-dimensional structure and biological activities of the reaction molecules. So, Hansch and Free-Wilson methods are called a two-dimensional quantitative structure-activity relationship. The two-dimensional quantitative structure-activity relationship had a great influence in the field of pharmaceutical chemistry. People's awareness of structure-activity relationships are from the traditional qualitative level to quantitative level. Quantitative structure-activity relationship reveals that the drug molecules and biological macromolecules combination mode to a certain extent. Under the guidance of Hansch method, people successfully designed the Norfloxacin and other Quinolone antibacterial drug.

Because of two-dimensional quantitative cannot accurately describe the relationship between the three-dimensional structure and biological activities, Scientists began to explore the feasibility of three-dimensional quantitative structure-activity relationship based on molecular conformation before and after the 1980s. Three-dimensional quantitative

2015 © The American Computational Science Society. All rights reserved.

structure-activity relationship is based on the introduction of the 3D structure information of drug molecules. This method indirectly reflects the interaction between drug molecules and biological macromolecules interaction process, compared with the two-dimensional quantitative structure-activity relationship has more explicit physical meaning and more abundant information. Since the 1980s, three-dimensional quantitative structure-activity relationship gradually replaced the position of two-dimensional quantitative structure-activity relationship, became one of the main methods of rational drug design based on the mechanism.

In this study, a new analysis strategy is used for safety evaluation of poison. The quantitative structure-activity relationship (QSAR) method is based on the chemical structure information and using quantitative parameters alone or combining with quantitative conventional parameters to construct the mathematical model, and using this model to forecast the toxicity of some unknown toxic compounds. QSAR can not only quickly establish a reliable forecasting mode, but also can reveal the damaging effect mechanism of the poison interacting with the body and provide the reference information of designing and synthesizing the safer and eco-friendly real green compounds [10-12]. The application of QSAR will open a new research method for toxicology. Gene expression programming (GEP) is a relatively new evolutionary computation algorithm, which is a kind of good data mining and modeling tools. Heuristic method (HM) has strong ability of function mining and can rapidly find the answer in complex problem. GEP and HM are used to evaluate the toxicological safety for nitrobenzene compound, and compare the stability of the model.

2. The main steps of QSAR

Mainly process of QSAR includes: collect and organize data sets, construction of molecular structure of compounds, calculation and selection of descriptor, establishment and validation of the model. QSAR study council had set up an international rule on the 2004 international conference that is OECD principles. The following are the condition of QSAR establishment: (1) A clear response values; (2) A determined algorithm; (3) A clear application domain; (4) Goodness of fit, robustness and predictive methods are appropriate; (5) The interpretation of mechanism.

2.1. Collection of data

The reliable data should not be small and must possess representativeness that can ensure the accurate of QSAR model. Usually data are from the thesis in the magazine and professional database. The data from the literature requests from the same laboratory researchers. Professional databases:

(1) http://toxnet.nlm.nih.gov/cpdb/cpdb.html.
(2) http://www.il-eco.uft.uni-bremen.de/login.php.
(3) http://www.safe.nite.go.jp/english/.

Sometimes, when the activity of a data set (nature) values across several orders of magnitude, requires the transformation of the data, usually convert the original values to the form of log10.

2.2. Calculation and selection of descriptor

Compounds are built molecular structural formula in Chemdraw software. Firstly, the geometry optimization is in the Hyperchem software, the calculation of molecular mechanics is with MM+. The Semi-empirical AM1 method can get more accurate optimization. New data will be further precisely calculated in the MOPAC [13]. Then, the HIN files generated by geometrical optimization, the MNO files generated by MOPAC calculation and the toxicity values will be passed in CODESSA software to calculate the optimum molecular descriptors. There are five groups of descriptors: Constitutional descriptors, Geometrical descriptors, Electrostatic descriptors, Thermodynamic descriptors and Topological descriptors, nearly 1500 kinds of descriptors.

2.3. Heuristic Method (HM)

HM can search all completely calculated descriptors, ultimately selected the best combination of descriptor. Before execution, HM pre filter the descriptors, well. The first is for descriptor linear restrictions [14]. In HM algorithm, if the correlation coefficient between any two descriptors >0.80, the final selection will only choose one of them. Second to eliminate some descriptors, mainly is the parameters of partial compounds not have. T test value is less than a defined. HM will order descriptors in descending according to the correlation coefficient of parameter model. Each introduction descriptor is the biggest of all response correlation coefficient values. The performance of the model is mainly verified by R^2, Q^2_{CV} and S^2. When Q^2_{CV} increment is less than 0.02, the selection process terminated. Q^2_{CV} is used to measure the reliability and robustness of QSAR model.

2.4. Gene expression programming (GEP)

GEP, a non-parametric method, is invented by Ferreira in 1999 [15], and it is an extension of genetic programming (GP). Gene expression programming is a genotype/phenotype system, in which a mathematical function defined as a linear chromosome with fixed length. The chromosomes are composed with multi-genes which are structurally organized in a head and a tail. The head of gene contains both function and terminal symbols, whereas the tail is terminal symbols only. GEP is individual chromosomes. Each chromosome is made up of one or more genes. Each gene is a fixed length string, divided into the head and tail. Head can also contain a terminator, and tail contains only the end operator. Terminators are often constant. For a chromosome, the same length of each gene is fixed, so the length of each chromosome is fixed. The length of the body color is determined by

2015 © The American Computational Science Society. All rights reserved.

the need to solve specific problems, but once we know head length, tail length can be calculated by formula. Suppose the head length of H, tail length of T, the relationship between t and h as: T = H*(n-1) + 1. N represents the function collection of input parameters that are the most number of parameters in the function. The function of chromosomes is subjected to modification by means of mutation, transposition, etc, to find a best solution in the form of expression trees (ETs). The genetic code of GEP is very simple: a one-to-one relationship between the symbols of chromosome and the nodes they represent in the trees. Operation process of GEP method is first to create a random initial population, the population of each individual translated into K a expression, and calculate the fitness value to assess individual fitness. If there are not termination conditions, it will update population by genetic operations [15].

This process has been circulating, until meet the termination conditions. GEP includes single point of restructuring, recombination, mutation, insertion, genetic recombination and other genetic operations that have a specific meaning. The real answer from GEP is an expression, namely individual phenotype. The expression is obtained by chromosome. Despite the length of the chromosome is fixed, but due to the special encoding, expression length can be different. GEP genetic operators in addition to the selection operator and mutation operator of genetic algorithm, but also increased the conversion and insertion operator and restructuring grates. In order to keep the decode expression tree completely, the genetic operations need to keep the length of the gene and the tails end up with operators.

2.5. The model validation

In QSAR studies, it is very necessary to validate the performance of the model. A stable QSAR model should have good fitting ability, high robustness and ability to predict. Internal validation and external validation are the common method for model validation [16,17].

Internal validation validates the goodness-of-fit and robustness of QSAR model. The greater of the determination coefficient (R^2), the model fitting ability is the better. Root Mean Squares Error (RMSE) is also used to determine the model performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{\exp} - y_i^{cal}\right)^2}{\sum_{i=1}^{n}\left(y_i^{\exp} - \bar{y}^{\exp}\right)^2}$$

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{\exp} - y_i^{cal}\right)^2}{n}}$$

$y_i^{\exp}$ is the experimental value of compounds; y_i^{cal} is the calculated value of compounds; $\bar{y}^{\exp}$ is the average of experimental value of compounds for training set; n is the number of training set.

For the training set, usually adopt the interactive test to verify the model's stability and reliability, in order to reduce the possibility of excessive fitting. Interactive authentication method: leave-one-out (LOO), n-fold cross validation and bootstrapping. LOO involves all of the training set of data, so it also is the most commonly validation method. Usually adopt correlation coefficient and root mean square error (RMSECV) of cross validation to determine the robustness of the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{\exp} - y_i^{calcv}\right)^2}{\sum_{i=1}^{n}\left(y_i^{\exp} - \bar{y}^{\exp}\right)^2}$$

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{\exp} - y_i^{calcv}\right)^2}{n}}$$

$y_i^{\exp}$ is the experimental value of compounds; y_i^{calcv} is the calculated value of cross validation of compounds; $\bar{y}^{\exp}$ is the average of experimental value of compounds for training set; n is the number of training set.

External validation is mainly used to measure model of external prediction ability. The most widely measures of the parameters are the decision coefficient (R^2) and the root mean square prediction error (RMSEP) of test set.

$$R^2 = 1 - \frac{\sum_{j=1}^{k}\left(y_j^{\exp} - y_j^{cal}\right)^2}{\sum_{j=1}^{k}\left(y_j^{\exp} - \bar{y}^{\exp}\right)^2}$$

$$RMSEP = \sqrt{\frac{\sum_{j=1}^{n}\left(y_j^{\exp} - y_j^{cal}\right)^2}{k}}$$

$y_j^{\exp}$ is the experimental value of compounds; y_j^{cal} is the calculated value of compounds; $\bar{y}^{\exp}$ is the average of experimental value of compounds for test set; n is the number of test set.

3. Conclusion

With the continuous development of the computing simulation method, comprehensive development of molecular structure descriptors calculation method and updating of hemometrics statistical simulation methods, QSAR develops from the initial linear MLR to ANN, SVM, LSSVM nonlinear simulation until now. QSAR has been successfully applied in various research fields, such as environmental science, chemistry, biology, drug design, material, etc. Research content extends from the original simple log P to the various physical and chemical properties of compounds, various toxic (poisonous to the cells and

2015 © The American Computational Science Society. All rights reserved.

ecological toxicity), carcinogenic teratogenicity, degradability, and various kinds of biological effect index. QSAR research results have also been used to compound security risk evaluation. QSAR model can be used to predict the activity of untested compounds or the physical and chemical properties. On the other hand, QSAR can reveal the structural factors which affect the properties of security risk of compounds, and further better understand the mechanism of action of the toxicity (ecotoxicity, cytotoxicity) and biodegradation and provide the reference information of designing and synthesizing the safer and eco-friendly real green compounds.

References

[1] J.A. Lopez, A.J. Brennan, J.C. Whisstock, I. Voskoboinik, J.A. Trapani, Trends Immunol., 33 (2012) 406-412.

[2] R.H.P. Law, N. Lukoyanova, I. Voskoboinik, T.T. Caradoc-Davies, K. Baran, M.A. Dunstone, M.E. D'Angelo, E.V. Orlova, F. Coulibaly, S. Verschoor, K.A. Browne, A. Ciccone, M.J. Kuiper, P.I. Bird, J.A. Trapani, H.R. Saibil, J.C. Whisstock, Nature, 468 (2010) 447-451.

[3] O. Naneh, A.B. Zavec, D. Pahovnik, E. Žagar, R.J. Gilbert, I. Križaj, G. Anderluh, J. Immunol. Methods, 426 (2015) 19–28.

[4] G.E. Janka, Annu. Rev. Med., 63 (2012) 233−246.

[5] M. Barry, R.C. Bleackley, Nat. Rev. Immunol., 2 (2002) 401-409.

[6] J.L. Veale, L.W. Liang, Q. Zhang, D.W. Gjertson, Z. Du, E.W. Bloomquist, J. Jia, L. Qian, A.H. Wilkinson, G.M. Danovitch, P.T. Pham, J.T. Rosenthal, C.R. Lassman, J. Braun, E.F. Reed, H.A. Gritsch, Hum. Immunol., 67 (2006) 777-786.

[7] M.C. Fantini, C. Becker, R. Kiesslich, M.F. Nuerath, Nat. Clin. Pract. Gastroenterol. Hepatol., 3 (2006) 633−644.

[8] N. Adhikari, A.K. Halder, A. Saha, K.D. Saha, T. Jha, Toxicol. In Vitro, 29 (2015) 1392-1404.

[9] L. Liu, X.R. Han, H.Z. Si, L.H. Cui, J. Comput. Sci. Eng., 12 (2014) 438-444.

[10] V.M. Patil, S. Gupta, S. Samanta, Lett. Drug Des. Discov., 7 (2010) 139-148.

[11] A.K. Bandyopadhyaya, J. Johnsamuel, A.S. Al-Madhoun, S. Eriksson, W. Tjarks, Bioorg. Med. Chem., 13(2005) 1681-1689.

[12] M. Uysala, H. Tanyildizi, Constr. Build Mater., 27 (2012)404-414.

[13] J.A. Spicer, G. Lena, D.M. Lyons, K.M. Huttunen, C.K. Miller, P.D. O'Connor, M. Bull, N. Helsby, S.M. Jamieson, W.A. Denny, A. Ciccone, K.A. Browne, J.A. Lopez, J. Rudd-Schmidt, I. Voskoboinik, J.A. Trapani, J. Med. Chem., 56 (2013) 9542-9555.

[14] J.P.P. Stewart, MOPAC 6.0, QCPE, No. 455, Quantum Chemistry Program Exchange, Indiana University, Bloomington, IN, 1989.

[15] A.R. Katritzky, V.S. Lobanov, M. Karelson, R. Murugan, M.P. Grendze, J.E. Toomey, Rev. Roum Chim., 41(1996) 851-868.

[16] J.S. Wright, D.J. Carpenter, D.J. McKay, K. Ingold, E. J. Am. Chem. Soc., 119 (1997) 4245-4252.

[17] H.Z. Si, J.G. Zhao, L.H. Cui, N. Lian, H.L. Feng, Y.B. Duan, Z.D. Hu, Chem. Biol. Drug Des., 78 (2011) 370–377.

2015 © The American Computational Science Society. All rights reserved.