Prediction of Retention Times for a Large Set of Pesticides Based on Improved Gene Expression Programming

Zhang Kejun
College of Computer Science and technology
Zhejiang University
Hangzhou, Zhejiang 310027 China
channy@zju.edu.cn

Sun Shouqian
College of Computer Science and technology
Zhejiang University
Hangzhou, Zhejiang 310027 China
ssq@zju.edu.cn

Si Hongzong
Institute for Computational Science and Engineering
Qingdao University
Qingdao, Shandong 266071, China
sihz03@126.com

ABSTRACT
The purpose of the paper is to present a novel way to building Quantitative structure-retention relationship (QSRR) models. Studies was reported for predicting the retention times (RTs) of 110 pesticides which were detected by gas chromatography (GC) with mass selective detector (MSD). Chemical descriptors were calculated from the molecular structure of pesticides and the QSRR models of RTs with descriptors was built using the heuristic method (HM) and Improved Gene Expression Programming (IGEP), respectively. The obtained linear model of HM had a correlation coefficient R^2 = 0.913, with a root mean square error (RMS) S^2 of 0.0387 for the training set, while R2 =0.907， and RMS =0.0408 for the test set. The nonlinear model by IGEP gave better results: for the training set R^2 = 0.971, S^2 = 0.0176 and for the test set R^2 =0.951, S^2 =0.0267. The prediction results from nonlinear model are in agreement with the experimental values The QSRR model also reveals that the gas chromatographic RTs are associated with physicochemical property of pesticides.

Categories and Subject Descriptors： I.2.1 Applications and Expert Systems

General Terms： Algorithms

Keywords
Improved Gene Expression Programming; Heuristic Method; Quantitative structure-retention relationship; Retention times

1. INTRODUCTION
Pesticides are necessary and essential in agricultural production. With their use, not only to raise the agricultural produce and ensure farms' profits, but also the risk of residues in the food consumed is present. For this reason, governments and international organizations have published a list of pesticides and their tolerances or maximum residues limits standards [1, 2].

In the present study, the aim was to explore the retention behavior of a series of pesticides in GC-MSD, to establish a new QSRR model and to confirm the possibility of predicting retention behavior of diverse derivatives. The results showed that the models obtained were satisfactory and the Improved Gene Expression Programming (IGEP) had a higher predicting power than HM in dealing with the non-linear questions. The structural factors affecting the compounds' retention behavior were also investigated.

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-130-9/08/07.

1.1 The heuristic method
Once molecular descriptors are generated, the heuristic method [3] in CODESSA was used to accomplish the pre-selection of the descriptors and build the linear model. Its advantages are the high speed and no software restriction on the size of the data set. Heuristic method can either quickly give a good estimation about what quality of correlation to expect from the

data, or derive several best regression models. All two-parameter regression models with remaining descriptors are subsequently developed and ranked by the regression correlation coefficient R^2. A stepwise addition of further descriptor scales is performed to find the best multi-parameter regression models with the optimum values of statistical criteria (highest values of R^2, means error and the F-value).

1.2 Gene Expression Programming
The classic GEP[7,8,9] algorithm considered insufficiently in the mutation operator result in performance insufficiency in keeping the gene multiplicity even appears the precocious phenomenon. It usually solved by initializing again the population which caused the algorithm efficiency to reduce. The author proposed a dynamic mutation operator algorithm to enhance the algorithm efficiency. The main purpose of dynamic mutation is: the number of mutations is decided by the number of genes in the chromosome.

Step1: Random producing a number RA between 0 and 1, if the RA greater than p_m (mutation rate), go to step2, else, exit.

Step2: Random producing a number RB between 0 and p_g (population size), and RB decide which chromosome will be mutated, go to Step 3.

Step3: Random producing c_p numbers between 0 and g_l (length of gene), where, c_p is the number of genes in the chromosome RB. The number is produced to be located the mutation bits of the chromosome.

2. EXPERIMENTAL
A complete list of the compounds name and their corresponding retention time (Log tR) for 110 pesticides was omitted. All compounds were collected from Chinese Academy of Inspection and Quarantine. The data set was divided into two subsets in HM and IGEP: a training set of 88 compounds and a test set of 22 compounds. The training set was used to build the HM and IGEP

model and the test sets was used to evaluate its prediction ability of both models.

To obtain a QSPR model, compounds are often represented by the molecular descriptors. The calculation process of the molecular descriptors is described as below: All molecules were drawn into Hyperchem, and pre-optimized using MM+ molecular mechanics force field. A more precise optimization was done with semi-empirical AM1 [10] method in MOPAC [11]. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient was 0.001. The MOPAC output files were used by the CODESSA program [12,13] to calculate five classes of descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.); topological (Wiener index, Randic indices, Kier–Hall shape indices, etc.); geometrical (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.); and quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.)[12]

3. RESULTS AND DISCUSSION

Total 526 descriptors were calculated by the CODESSA program for all the compounds. To select the set of descriptors that are most relevant to the pesticide process of organic pollutants, to show the affecting degree of different descriptors for pesticide process and well understand the accumulation mechanism of chemicals in organism, the linear models with the number of variables from 1 to 9 were built.

Also, It used CPSS [14] which based on the IGEP to model this function because it allows the easy optimization of intermediate solutions and the easy testing of the evolved models against a test set. In one run a very good solution with a CC/ R^2 of training set was 0.971.

The experiment shows the prediction results by IGEP (either training set or testing set) are better than HM. Therefore, IGEP as a non-linear method has good generalized performance (Table 1).

Table 1. Results of CC and R by HM, SVM and IGEP

Methods	Training set		Test set	
	R2	RMS	R2	RMS
HM	0.913	0.0387	0.907	0.0408
IGEP	0.971	0.0176	0.951	0.0267

4. CONCLUSIONS

QSRR models for the prediction of retention time of pesticides compounds using the heuristic method and Improved Gene Expression Programming based on descriptors calculated from molecular structure alone have been developed. Satisfactory results were obtained with the proposed method. The proposed linear model could identify and provide some insight into what structural features are related to retention time of these compounds. Additionally, nonlinear IGEP model based on the same sets of descriptors showed better predictive ability. The good predictive ability of the models allows us to estimate retention indices for similar compounds in cases where retention values are not readily available.

5. ACKNOWLEDGMENTS

This research was supported by National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 20050335096.

6. REFERENCES

[1] Analytical methods for Pesticide residues in foodstuff, 6thed., General Inspectorate for Health Protect, Ministry of Health, Welfare and Spot, Amsterdam, The Netherlands,1996. 1

[2] Ock Kyoung Chun,hee Gon Kang, Estimation of risks of pesticide exposure, by food intake, to Koreans. Food and chemical Toxicology 41:1063-1076(2003). 2

[3] Katritzky, A.R.; Lobanov, V.S.; Karelson, M. Reference Manual, Version 2.0, 1994.

[4] Ferreira C, Gene expression programming: A new adaptive algorithm for solving problems., Complex Systems, 13(2001) 87-129.

[5] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MIT, 1975

[6] D.E. Goldberg, Genetic algorithms in search, optimization, and machine learning, Addison-Wesley, Reading, MA, 1989

[7] Koza J R. Genetic Programming: On the Programming of computers by Means of Natural Selection. Cambridge, MA:MIT press,1992

[8] Ferreira, C., 2002. Gene Expression Programming in Problem Solving. in Roy, R., S. Ovaska, T. Furuhashi, and F. Hoffman, Eds., Soft Computing and Industry-Recent Applications, Springer-Verlag, pp. 635-654.

[9] M. Mitchell, An Introduction to Genetic Algorithms, in Complex Adaptive Systems. (MIT Press, 1996).(OK)

[10] Dewar, M.J.S., Zoebisch, E.G., Healy, E.F., Stewart, J.J.P., Development and use of quantum molecular models. 75. Comparative tests of theoretical procedures for studying chemical reactions. J. Am. Chem. Soc. 107: 3898–3902(1985).

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

[12] Katritzky, A.R., Lobanov, V.S., Karelson, M., Comprehensive descriptors for structural and statistical analysis, Reference Manual, Version 2.0. 1994.

[13] atritzky, A.R., Lobanov, V.S., Karelson, M.,.QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chem. Soc. Rev. 24:279–287(1995).

[14] Zhang Kejun, Study of Improved Gene Expression Programming for Solving Inverse problem, in 2006 Dissertation of Jiangxi University of Science and Technology, Chap 4.