Chinese Chemical Letters Vol. 14, No. 9, pp 987– 990, 2003
http://www.imm.ac.cn/journal/ccl.html

Application of Genetic Programming in Predicting Infinite Dilution Activity Coefficients of Organic Compounds in Water

Yi Lin CAO*, Huan Ying LI

College of Chemistry and Environmental Science, Henan Normal University, Xinxiang 453002

Abstract: In this paper, we calculated 37 structural descriptors of 174 organic compounds. The 154 molecules were used to derive quantitative structure – infinite dilution activity coefficient relationship by genetic programming, the other 20 compounds were used to test the model. The result showed that molecular partition property and three-dimensional structural descriptors have significant influence on the infinite dilution activity coefficients.

Keywords: Infinite dilution activity coefficients, genetic programming.

The infinite dilution activity coefficient (γ^{∞}) is of both theoretical and practical interest, which provides insight into the kinds of chemical and physical forces experienced between solvent and solute molecules. Its value is important for the development of new thermodynamic model, the design of separation and extraction processes and the selection of solvents for rectification and extraction1. The value of γ^{∞} is commonly obtained by extrapolating data to infinite dilution concentration. Various methods for measuring γ^{∞} have been described in literature2, but each method has some specific disadvantages. For reasons of time, cost, safety and technical availability, it is necessary to establish a quantitative relationship between molecular structures and γ^{∞} values.

Molecular structures descriptors

It is a fundamental tenet of chemistry that the structures of compounds determine their chemical, physical and biological properties. Here, we employed 37 relevant structural descriptors of 174 organic molecules respectively.

Theory

Genetic Programming (GP) developed by Koza is a random optimization method based on the concepts of biological evolution and natural selection. It randomly creates computer program that represents the solution, the size of program changes with the complexity of problems to be solved. It usually uses 3 steps to solve the problems:

* E-mail: hxxcyl@eyou.com

1. Generate randomly an initial population (the form of a binary tree) which is stratification computer programs including the functions set and terminators.
2. Iteratively perform the following steps until the termination criteria have been satisfied.

(1). To compute the fitness of each program in the population according to how well it solves the problem.

(2). To create a new population of computer programs by applying copy, crossover and mutation.

3. The best computer program, the best-so-far solution, is designated as the result of GP. Detailed information about GP can be found in literature[3].

Results and Discussion

The data of γ^{∞} were from literature[4]. The 37 calculated descriptors for each molecule include partition property, constitutional descriptors, charge descriptors, 2D autocorrelation indices, Randic molecular profile indices, geometrical descriptors and 3D-MoRSE descriptors, such as LogP, Sv , Se, qpos, Q_2, GATS8P, GATS9M, SP01, SP35, G_2, J_{3D}, Mor11v and Mor21v *etc*. Not all of these descriptors were correlated with γ^{∞}, so we employed GP techniques to select among these descriptors and then correlate them with γ^{∞}. First, the compounds were divided into two sets, one is called training set including 154 molecules and the other is called test set including 20 molecules. To form initial population, we employed the function set {+, –, x, /, √ } and terminators {LogP, SP03, Se, Morse11v *etc*. and some random numbers between [-1, +1]}. Mean fitness, defined as the squared error measurement, was used to direct the evolution process. To run the program, several parameters were set as follows: population size is 300, program size was set between 80 and 256, crossover probability is 40%, mutation probability is 95%, and copy probability is 3%. The program was stopped until 113614 generations passed. After simplification and optimization, an equation with 10 parameters was derived as follows.

$$A = \log P - 5Q_2$$

$$B = [(Mor21v + \log P)^2 - 1.1813 Mor11v] / Se^2$$

$$C = 4/(\log P * SP03 * 1.4625 * Se^2)$$

$$D = \{4.7252[(Mor21v + \log P)^2 - Mor11v] + SP04 - SP18 + 0.3514\} / 2.1917 Se^2 - 4$$

$$E = 4[0.7126(Mor21v + \log P)^2 / \log P]^2 / Se^2$$

$$F = \{2[2(1 + sp05) / \log P - 0.2193]\}^{1/2} / Se^3$$

$$G = 1.4625 - 4 / J_{3D}^2 - 54.1089 J_{3D}^4$$

$$\log \gamma^{\infty} = A + \{[C-D+E+B+F+F/(E-D+B+F)]^{1/2} + G\}^{1/2}$$

N=154 R=0.9383 SD=0.5092

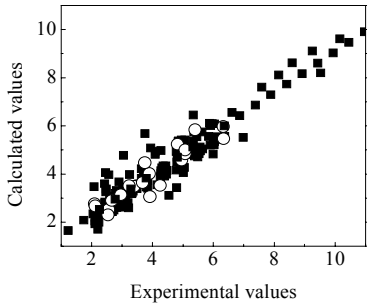
Here R is correlation coefficient, SD is standard deviation. The sign A through G denote the entries in the last equation. logP is partition coefficients between *n*-octanol and water, Se is sum of atomic Sanderson electronegtivities, Q_2 is total squared charge, SP03, SP04, SP05 and SP18 are a sequence of global molecular shape profiles, J_{3D} is Balaban distance connectivity index, Mor11v and Mor21v are 3D-MoRSE descriptors.

Detailed information of these descriptors can be found in literature[5]. γ^{∞} values of 20 compounds left out were predicted using the above equation. The SD of prediction is 0.4272 in log unit and comparison between calculations and experiments was shown in **Figure 1**. Here the solid squares(■) denote the training set and the hollow circles(○) denote the test set. The prediction results were listed in **Table 1**. It can be seen that the calculations and predictions (Pre.) were fitted well with the experiments (Exp.).

Table 1 Comparison between predictions and experiments of 20 compounds

Compounds	Exp.	Pre.	Compounds	Exp.	Pre.
2,2-dimethyl-1-butanol	2.5635	2.5168	1-butanethiol	3.9206	3.0450
2-methyl-2-butene	4.2552	3.5187	3-hexanol	2.5441	2.2881
2,4-dimethyl-3-pentanol	2.9619	3.1383	1-octene	5.4031	5.8299
2,4-dimethyl-2,4-nonanediol	3.3944	3.4798	2-methylhexane	6.3304	5.9586
chlorodifluoromethane	3.2380	3.4979	*n*-octane	4.9585	4.5714
2,4-dimethyl-2-pentanol	2.6758	2.8979	1-bromoprntane	4.8208	5.2348
2,2-dimethyl-3-pentanol	3.8926	4.0046	pentanoic acid	2.1038	2.7645
1,4-cyclohexadiene	3.6776	3.6303	butylethylamine	2.1139	2.6593
4-ethylcyclohexane	5.0828	4.8316	2-methylhexene	6.3404	5.4635
nitrotrichloromethane	3.7543	4.4620	1,6-heptadiene	5.0791	4.9892

Figure 1 Scatter plot of calculations *vs* experiments in log unit



The parameters selected into the equation can be divided into two types: one is the partition property, such as LogP, the other is one which includes several 3D structural descriptors that express the molecular size, shape, branch and charge characteristic, and both of them are often of great importance for prediction and description of molecular properties.

Moreover, the GP method has two special characteristics, it can find the optional variable combination. In this study, 10 descriptors were selected from 37 descriptors. The other is that GP can automatically determine the linear or nonlinear model. In the study here, it derived a nonlinear equation.

Acknowledgments

This work was supported by Natural Science Foundation of Henan Province (No.532221).

References

1. L. Dallinga, M. Schiller, J. Gmehling, *J. Chem. Eng. Data*, **1993**, *38* (1), 147.
2. R. A. Katrizky, S. V. Lobanov, *Chem. Soc. Rev.*, **1995**, *24* (5), 279.
3. J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, The MIT Press, Cambridge, **1992**.
4. S. R. Sherman, D. B. Trampe, D. M. Bush, M. Schiller, C.A. Eckert, A. J. Dallas, J. J. Li, P. W. Carr, *Ind. Eng. Chem. Res.*, **1996**, *35* (4), 1044.
5. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, WILEY-VCH Verlag GmbH, Weinheim, **2000**.

Received 18 september, 2002