Chromatography
Separation Techniques

Zhang et al., J Chromatogr Sep Tech 2015, 6:7
DOI: 10.4172/2157-7064.1000306

Research Article Open Access

# Study on Quantitative Structure-Retention Relationships (QSRR) for Oxygen-Containing Organic Compounds Based on Gene Expression Programming (GEP)

**Zhang X[1,2]*, Shi L[2], Ding L[2], Sun Z[2], Song L[2], Qu H[3] and Sun T[1]**

[1]College of Sciences, Northeastern University, 110004 Shenyang, Liaoning, People's Republic of China
[2]Liaoning Key Laboratory of Petrochemical Engineering, Liaoning Shihua University, 113001 Fushun, Liaoning, People's Republic of China
[3]Research Institute of PetroChina, Fushun Petrochemical Company, 113004 Fushun, Liaoning, China

## Abstract

Gene Expression Programming (GEP) is a novel genetic algorithm, a highly effective, stable random searching method. We take GEP to make models of Quantitative Structure-Retention Relationship (QSRR) for a series of oxygen-containing organic compounds of GC retention index, and compare the predictive results with Artificial Neural Network (ANN) and Multiple Linear Regression (MLR). The correlation coefficient on OV-1 column is 0.9919, 0.9891 and 0.9911 for GEP, ANN and MLR respectively, on SE-54 column is 0.9955, 0.9892, and 0.9917. It is shown that the predicted results by GEP are in good agreement with experimental ones, better than those of ANN and MLR.

**Keywords:** Gene Expression Programming (GEP); Oxygen-containing organic compounds; Artificial Neural Network (ANN); Quantitative Structure-Retention Relationship (QSRR)

## Introduction

Chromatography in itself is not an accurate analytical technique, but rather a separation one. The identification of oxygen-containing organic compounds can be made with the method of gas chromatographic peak in comparison with that of a standard sample of each compound. Because samples of pure compounds are not always available, it is important to develop QSRR that can efficiently predict retention parameters by using theoretical descriptors computed from chemical structure.

Quantitative Structure-Retention Relationships (QSRR) [1] establish the relationship between a chemical structure and its chromatographic retention value, which has been demonstrated to be a powerful tool for the investigation of chromatographic parameters. The main advantage of QSRR is the ability to distinguish in quantitative theoretical terms, packing materials of different chemical nature of the organic ligand and/or organic or inorganic support [2], Furthermore, it can be of valuable assistance in the prognosis of the behavior of new molecules, even before they are actually synthesized [3].

An important property that has been extensively studied in QSRR is the chromatographic retention index. The retention index is a generally accepted type of data used for the identification of chemical compounds by gas chromatography. A retention index is a continuous quantitative variable that relates the retention of a solute to the retention of a set of standard compounds. Retention indices are much less dependent on experimental factors (e.g., Temperature, flow, column, length etc.) than retention times. While Kovats retention indice [4] have linear collerations with column temuprature. And they were obtained by the logarithmic interpolation method.

QSRR on the Kovats retention indices have been reported for different types of organic compounds. The Kovats retention index is the most popular dependent variable in QSRR studies because of its reproducibility and accuracy. In many cases, the precision and accuracy of the QSRR models are not sufficient for identification purposes; still the models are useful to elucidate retention mechanisms, to optimize the separation of complex mixtures or to prepare experimental designs.

Topological descriptors computed on the basis of molecular graph are easy to be calculated with present computing facilities. Due to the simplicity and efficiency of graph-theoretical approaches, we take novel polarizability effect index (PEI), odd-even index (OEI), the sum eight values $X_{1CH}$ of every C-H bond adjacency matrix $S_{x1CH}$.

An interesting and increasing application of QSRR is to test various chemometric methods from multiple linear regression (MLR) methods to Artifical neural network (ANN) methods. Multiple linear regression (MLR) is without doubt the most frequently applied technique in building QSRR models.

Gene Expression Programming (GEP) is a new evolutionary algorithm that evolves from computer programs (they can take many forms: mathematical expressions, neural networks, decision trees, polynomial constructs, logical expressions, and so on). The computer programs of GEP, irrespective of their complexity, are all encoded in liner chromosomes. Then the liner chromosomes are expressed or translated into expression trees (the branched structures). Thus, in GEP, the genotype (the liner chromosomes) and the phenotype (the expression trees) are different entities (both structurally and functionally), and because of this apparently trivial fact, this new evolutionary system can finally make a difference, successfully assisting researchers in the design of robust and accurate computer models [5].

The aim of the present research is to develop a general model capable of predicting the gas chromatographic retention index of oxygen-containing organic compounds based on GEP, compared with the result predicted by traditional linear MLR and another powerful non-linear ANN method.

***Corresponding author:** Zhang Xiaotong, College of Sciences, Northeastern University, 110004 Shenyang, Liaoning, People's Republic of China, Tel: +86-024-56860658; Fax: +86-024-56860658; E-mail: xt_zhang2015@163.com

**Received** November 19, 2015; **Accepted** November 28, 2015; **Published** December 10, 2015

**Citation:** Zhang X, Shi L, Ding L, Sun Z, Song L, et al. (2015) Study on Quantitative Structure-Retention Relationships (QSRR) for Oxygen-Containing Organic Compounds Based on Gene Expression Programming (GEP). J Chromatogr Sep Tech 6: 306. doi:10.4172/2157-7064.1000306

## Materials and Methods

### Data set

The Kovats retention index of 91 molecules (include esters, ketones, and alcohols) taken from reference [6] were presented in Table 1. Kovat's retention index of all compounds was obtained under the same conditions on two stationary phases: OV-1 (dimethylpolysiloxane) and SE-54(5% phenyl -95% dimethylpolysiloxane) 74 molecules were used as training set for model generation and 17 molecules were used as test set for model prediction. The corresponding experimental and predicted values of the RI for all the molecules studied in this work are shown in Table 1.

### GEP theory

Gene Expression Programming (GEP) was first proposed formally by Candida Ferreira in 2001. It was an elegant and efficient solution to expression-mutation problems. GEP, which is an extraordinarily powerful tool, is a subset of Genetic Algorithms, except it uses genomes whose strings of numbers represent symbols. GEP-an evolutionary algorithm inherits both the evolutionary simplicity of Genetic Algorithms (GA) and the expressional power in Genetic Programming (GP) by utilizing a genotype/phenotype representation system. The string of symbols can further represent equations, grammars, or logical mappings.

Ferreira [5] proposes the use of a set of genetic operators: Replication, Mutation, IS Transposition, RIS Transposition, Gene Transposition, 1-Point Recombination, 2-Point Recombination, Gene Recombination. As Ferreira comments, the advantages of a Genetic Representation like the one in GEP are simple entities: linear, compact, relatively small, easy to manipulate genetically. The genetic operators applied to them are less restricted than those used in GP [5].

Fortunately for us, in GEP, thanks to the simple rules that determine the structure of expression trees and their interactions, it is possible to infer immediately the phenotype given the sequence of a gene. It is easy for a computer program to follow these three rules while performing mutations, and it never has to check whether the resulting expression has valid syntax. By allowing a broad range of mutations, the process can efficiently explore a high dimensional space, and the expressions can change in size as functions are replaced by terminals and terminals by functions.

GEP are evolutionary tools inspired in the Darwinian principle of natural selection and survival of the fittest individual and uses populations of candidate solutions to a given problem in order to evolve new ones. These methods use an initial random population and apply genetic operators to this population until the algorithm finds an individual that satisfies some termination criteria. The evolving populations undergo selective pressure and their individuals are submitted to genetic operators.

**Gene representation:** GEP genes are composed of a head and a tail. The head contains symbols that represent both functions (elements from the function set *F*) and terminals (elements from the terminal set *T*), whereas the tail contains only terminals. Therefore, two different alphabets occur at different regions within a gene. For each problem, the length of the head *h* is chosen, whereas the length of the tail *t* is a function of *h* and the number of arguments of the function with the most arguments *n*, and is evaluated by the equation:

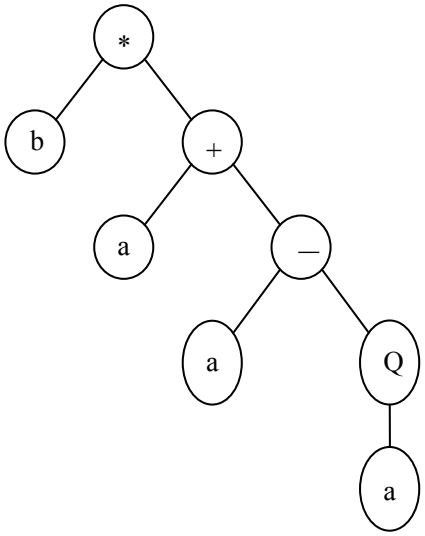
$$t=h\,(n\text{-}1)+1 \quad (1)$$

Consider a gene composed of {Q, *, /, -, +, a, b}. In this case *n*=2. For instance, for *h*=15 and *t*=16, the length of the gene is 10+11=21. One such gene is shown below (the tail is shown in bold):

012345678901234567890123456789

*b+a-aQab+//+b+**babbabbbababbaaa** (2)

It codes for the following ET:




A K-expression can be mapped into an expression tree (ET) following a first-order procedure. A branch of the ET stops growing when the last node in this branch is a terminal. For example, the ET shown above corresponds to chromosome (2). In this case, the open reading frames end at position 7, whereas the gene ends at position 30.

Chromosomes in GEP are usually composed of more than one gene of equal length. For each problem or run, the number of genes, as well as the length of the head, is chosen. Each gene codes for a sub-ET and the sub-ETs can be linked by pre-defined rules forming a more complex multi-subunit ET.

**Selection method and genetic operators:** References [3] suggests there is no difference between different selection methods. It is strongly advised to use a simple elitism in any GEP implementation. The elitism means copying the best (or few best) individual to the offspring population without modifying them. GEP uses the well-known roulette-wheel method for selecting individuals.

In GEP, individuals were selected according to fitness by roulette wheel sampling coupled with the cloning of the best individual. The fitter the individual is, the higher the probability of leaving more offspring. Thus, during replication the genomes of the selected individuals are copied as many times as the outcome of the roulette. The roulette is spun as many times as there are individuals in the population, always maintaining the same population size.

GEP uses simple elitism of the best individual of a generation, preserving it for the next one. Replication is an operation that aims to preserve several good individuals of the current generation for the next one. In fact, this is a do-nothing probabilistic operation that takes place during selection (using the roulette-wheel method), and replicated individuals will be subjected to the action of the genetic operators. The mutation operator aims to introduce random modifications into a given chromosome. A particularity of this operator is that some integrity rules must be obeyed so as to avoid syntactically invalid individuals. In the head of a gene, both terminals and functions are permitted (except

| No. | Compounds | RIov-1$_{exp}$ | RIse-54$_{exp}$ | RIov-1$_{pre}$ | RIse-54$_{pre}$ | RIov-1$_{pre}$ | RIse-54$_{pre}$ | RIov-1$_{pre}$ | RIse-54$_{pre}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | MLR | | ANN | | GEP | |
| Training | | | | | | | | | |
| 1 | 3,3-Dimethyl-1-butanol | 778.77 | 763.63 | 776.89 | 788.80 | 782.22 | 801.60 | 760.36 | 761.76 |
| 2 | 3-Methyl-3-hexanol | 826.62 | 841.11 | 814.26 | 836.41 | 831.95 | 835.48 | 815.86 | 836.80 |
| 3 | 2,2,4-Trimethyl-3-pentanol | 881.49 | 894.00 | 847.18 | 861.53 | 885.53 | 899.84 | 835.29 | 853.68 |
| 4 | 4-Methyl-1-pentanol | 821.19 | 836.97 | 809.72 | 823.22 | 822.74 | 844.74 | 804.57 | 811.72 |
| 5 | 2-Pentanone | 666.34 | 687.79 | 671.52 | 690.74 | 673.76 | 678.33 | 677.08 | 703.33 |
| | Isopropyl acetate | 646.54 | 661.78 | 660.33 | 679.07 | 667.98 | 673.74 | 661.57 | 682.94 |
| 6 | Propyl formate | 605.79 | 623.60 | 618.75 | 632.11 | 615.20 | 626.31 | 625.90 | 634.54 |
| 7 | Isobutyl formate | 673.40 | 689.84 | 699.24 | 712.76 | 681.99 | 690.37 | 707.24 | 711.61 |
| 8 | 4-Ethyl-3-hexanol | 953.25 | 967.63 | 944.88 | 962.88 | 947.61 | 963.90 | 944.37 | 947.38 |
| 9 | Butyl formate | 707.64 | 725.53 | 711.12 | 725.78 | 699.74 | 724.78 | 724.97 | 736.01 |
| 10 | 2,4-Dimethyl-2-pentanol | 775.91 | 789.03 | 783.98 | 803.39 | 794.79 | 805.14 | 776.97 | 813.11 |
| 11 | 2-Hexanone | 767.93 | 790.03 | 769.55 | 788.32 | 761.54 | 781.05 | 774.87 | 806.38 |
| 12 | 1-Heptanol | 955.05 | 971.73 | 932.05 | 945.08 | 953.74 | 967.40 | 948.86 | 955.97 |
| 13 | Isobutyl propionate | 852.83 | 869.02 | 855.00 | 872.73 | 847.41 | 863.52 | 851.15 | 865.93 |
| 14 | 2-Ethyl hexanal | 934.65 | 954.71 | 936.68 | 953.13 | 942.20 | 942.37 | 944.85 | 952.54 |
| 15 | 2,2,4,4-Tetramethyl-3-pentanone | 900.00 | 914.09 | 873.03 | 882.66 | 903.85 | 909.81 | 859.99 | 882.02 |
| 16 | 3,3-Dimethyl-2-butanone | 693.05 | 711.58 | 698.23 | 715.29 | 682.14 | 702.27 | 690.34 | 714.12 |
| 17 | Butyl butyrate | 979.36 | 997.07 | 974.40 | 991.25 | 971.71 | 987.65 | 972.98 | 988.62 |
| 18 | 2-Methyl-3-hexanone | 819.95 | 838.42 | 830.93 | 849.58 | 821.44 | 832.20 | 832.23 | 853.74 |
| 19 | 2,2-Dimethyl-1-propanol | 657.34 | 670.46 | 643.91 | 656.76 | 661.46 | 673.92 | 614.17 | 619.31 |
| 20 | 2-Methyl-2-heptanol | 916.43 | 930.38 | 915.41 | 933.85 | 926.28 | 939.32 | 912.96 | 926.32 |
| 21 | Methyl propionate | 615.21 | 630.43 | 590.22 | 608.00 | 614.74 | 627.78 | 594.77 | 616.39 |
| 22 | Methyl isobutyrate | 670.97 | 686.58 | 671.39 | 690.34 | 666.38 | 682.36 | 665.89 | 686.90 |
| 23 | 3,6-Dimethyl-3-heptanol | 986.60 | 1000.00 | 991.41 | 1008.20 | 973.88 | 995.18 | 977.26 | 988.84 |
| 24 | 3-Methyl-2-butanone | 640.92 | 661.44 | 648.96 | 668.02 | 652.44 | 665.53 | 649.82 | 671.46 |
| 25 | 3-Methyl-1-butanol | 719.03 | 734.39 | 719.66 | 734.47 | 711.24 | 732.08 | 703.23 | 711.51 |
| 26 | 2,2-Dimethyl-3-pentanol | 805.63 | 818.97 | 792.82 | 809.97 | 796.39 | 806.42 | 776.29 | 803.69 |
| 27 | 2-Ethylbutyl acetate | 956.99 | 974.66 | 976.12 | 991.47 | 970.80 | 987.58 | 979.65 | 969.91 |
| 28 | Isobutyl isobutyrate | 900.00 | 915.56 | 938.84 | 954.97 | 915.96 | 919.04 | 928.49 | 932.01 |
| 29 | Methyl hexanoate | 907.01 | 925.46 | 895.42 | 912.68 | 901.95 | 917.79 | 895.05 | 914.34 |
| 30 | 6-Methyl-2-heptanol | 951.10 | 965.00 | 956.36 | 970.21 | 956.90 | 967.592 | 963.69 | 964.78 |
| 31 | 3-Heptanone | 865.79 | 886.89 | 853.08 | 871.55 | 853.70 | 867.21 | 862.98 | 887.18 |
| 32 | 2-Methyl pentanal | 742.38 | 762.95 | 782.17 | 796.35 | 748.25 | 768.33 | 784.17 | 793.87 |
| 33 | 3-Pentanone | 676.41 | 700.00 | 663.17 | 683.25 | 668.70 | 674.72 | 670.08 | 698.03 |
| 34 | Butyl isobutyrate | 938.55 | 954.26 | 937.89 | 954.15 | 938.62 | 946.85 | 937.12 | 950.04 |
| 35 | Ethyl hexanate | 982.90 | 1000.00 | 982.29 | 998.68 | 973.80 | 989.91 | 980.57 | 994.26 |
| 36 | 2-Methyl-2-pentanol | 717.57 | 731.39 | 717.23 | 739.35 | 695.81 | 696.02 | 711.67 | 741.92 |
| 37 | Pentyl acetate1 | 896.36 | 914.88 | 884.62 | 901.68 | 901.11 | 912.56 | 890.22 | 910.48 |
| 38 | 2-Ethyl-1-butanol | 825.94 | 841.00 | 818.16 | 834.20 | 814.70 | 831.34 | 806.85 | 828.43 |
| 39 | Propyl butyrate | 881.53 | 898.88 | 874.68 | 893.58 | 883.32 | 892.54 | 873.51 | 898.84 |
| 40 | 2-Octanone | 968.77 | 991.27 | 966.20 | 980.92 | 970.00 | 983.02 | 982.88 | 997.12 |
| 41 | 2,4-Dimethyl-3-pentanone | 779.01 | 795.28 | 789.16 | 806.77 | 779.97 | 788.80 | 787.96 | 819.42 |
| 42 | Ethyl isovalerate | 838.35 | 854.28 | 854.48 | 872.20 | 847.38 | 863.18 | 850.94 | 865.74 |
| 43 | | | | | | | | | |
| 44 | Butyl acetate | 796.18 | 814.16 | 785.22 | 803.81 | 796.99 | 823.92 | 787.28 | 813.77 |
| 45 | Methyl butyrate | 705.61 | 722.96 | 690.13 | 709.13 | 702.13 | 721.05 | 692.14 | 714.29 |
| 46 | 2-Methyl butanal | 636.32 | 657.70 | 623.15 | 643.91 | 632.64 | 659.32 | 633.06 | 659.06 |
| 47 | 5-Methyl-3-heptanol | 943.58 | 957.88 | 950.68 | 967.13 | 952.21 | 966.46 | 952.35 | 960.19 |
| 48 | 2-Pentanol | 682.66 | 700.00 | 687.43 | 707.18 | 680.20 | 702.12 | 680.42 | 707.20 |
| 49 | 4-Heptanol | 875.42 | 890.00 | 860.20 | 879.34 | 866.68 | 878.27 | 862.49 | 886.06 |
| 50 | 3-Methyl-2-butanol | 666.02 | 680.26 | 665.02 | 684.60 | 656.48 | 687.14 | 653.13 | 674.30 |
| 51 | 3-Methyl-2-pentanone | 734.76 | 754.92 | 752.44 | 771.85 | 745.13 | 755.42 | 754.10 | 786.74 |
| 52 | 4-Heptanone | 853.35 | 873.44 | 849.97 | 868.78 | 850.96 | 861.75 | 859.57 | 884.48 |
| 53 | 2-Methyl-2-hexanol | 817.33 | 831.38 | 820.98 | 841.95 | 831.84 | 838.71 | 816.22 | 843.30 |
| 54 | 2-Methyl-2-butanol | 626.20 | 640.33 | 632.83 | 655.60 | 616.68 | 658.45 | 629.46 | 657.68 |
| 55 | Ethyl butyrate | 784.04 | 784.04 | 779.76 | 799.69 | 790.90 | 802.39 | 777.71 | 807.41 |
| 56 | 2-Ethyl-4-methyl-1-pentanol | 972.00 | 972.00 | 988.87 | 999.82 | 967.29 | 979.45 | 996.57 | 978.56 |
| 57 | 3-Hexanone | 764.84 | 764.84 | 757.80 | 777.76 | 757.62 | 773.78 | 763.51 | 797.35 |
| 58 | 2,2,4-Trimethyl-1-pentanol | 930.00 | 930.00 | 929.19 | 937.52 | 928.37 | 942.34 | 925.75 | 909.71 |
| 59 | Isobutyl acetate | 757.65 | 757.65 | 757.87 | 775.33 | 757.92 | 771.30 | 755.04 | 769.79 |
| 60 | 1-Hexanol | 852.96 | 852.96 | 871.65 | 887.15 | 856.71 | 869.69 | 854.25 | 870.01 |
| 61 | 3-Ethyl-3-pentanol | 843.09 | 843.09 | 814.63 | 838.50 | 832.49 | 846.81 | 814.02 | 847.98 |
| 62 | 2,2-Dimethyl-pentanol | 867.57 | 867.57 | 852.42 | 864.52 | 859.41 | 882.94 | 840.44 | 846.05 |
| 63 | 4-Methyl-2-pentanol | 744.14 | 744.14 | 756.88 | 774.96 | 753.54 | 757.67 | 746.66 | 767.60 |
| 64 | Methyl isovalerate | 761.30 | 761.30 | 775.82 | 793.62 | 758.72 | 766.78 | 762.58 | 776.21 |
| 65 | 5-Methyl-3-hexanol | 838.15 | 838.15 | 843.71 | 861.74 | 848.59 | 862.99 | 841.32 | 861.91 |
| 66 | 2,2-Dimethyl-3-heptanone | 964.66 | 964.66 | 981.29 | 993.86 | 958.27 | 980.05 | 966.94 | 984.12 |
| 67 | 2,3-Dimethyl-3-pentanol | 823.66 | 823.66 | 803.38 | 826.32 | 813.98 | 837.36 | 798.54 | 823.59 |
| 68 | 2-Methyl-1-pentanol | 818.35 | 818.35 | 802.63 | 817.40 | 805.03 | 823.58 | 791.29 | 801.34 |
| 69 | Isobutyl alcohol | 611.31 | 626.00 | 722.87 | 745.99 | 674.91 | 682.21 | 710.77 | 742.57 |
| 70 | 2,4-Dimethyl-3-heptanol | 821.18 | 821.18 | 809.41 | 827.53 | 811.85 | 821.81 | 794.67 | 826.13 |
| 71 | 1-Butanol | 646.48 | 646.48 | 644.18 | 659.18 | 640.48 | 663.39 | 619.70 | 635.72 |
| 72 | 5-Methyl-2-hexanone | 836.53 | 836.53 | 847.63 | 863.78 | 835.65 | 868.52 | 854.67 | 870.33 |
| 73 | 5-Methyl-3-hexanone | 816.74 | 816.74 | 832.51 | 850.18 | 824.34 | 842.85 | 838.03 | 857.75 |
| 74 | 2-Heptanol | 885.57 | 885.57 | 879.65 | 897.16 | 888.20 | 902.17 | 882.22 | 901.90 |
| Test set | | | | | | | | | |
| 1 | Propyl acetate | 693.34 | 713.63 | 683.42 | 702.30 | 617.44 | 637.53 | 689.44 | 711.89 |
| 2 | Ethyl propionate | 694.19 | 711.16 | 684.77 | 704.71 | 618.47 | 639.56 | 685.73 | 710.38 |
| 3 | Butyl propionate | 891.40 | 909.12 | 875.79 | 894.30 | 886.56 | 899.39 | 876.32 | 900.71 |
| 4 | 4-Methyl-2-pentanone | 721.24 | 741.61 | 742.54 | 760.17 | 647.07 | 678.29 | 742.86 | 762.88 |
| 5 | 2-Heptanone | 868.70 | 891.01 | 866.39 | 883.51 | 820.26 | 869.43 | 877.86 | 899.02 |
| 6 | 3-Pentanol | 684.21 | 700.00 | 676.76 | 697.32 | 625.59 | 636.85 | 672.45 | 700.92 |
| 7 | 3-Methyl-1-pentanol | 828.82 | 845.00 | 831.83 | 846.72 | 769.50 | 787.37 | 823.22 | 841.33 |
| 8 | 4-Octanol | 975.50 | 990.22 | 956.58 | 973.54 | 960.82 | 979.67 | 962.16 | 980.55 |
| 9 | Propyl propionate | 792.58 | 809.79 | 764.89 | 784.53 | 713.60 | 724.93 | 771.67 | 802.09 |
| 10 | 1-Pentanol | 750.40 | 766.59 | 736.52 | 752.01 | 656.55 | 668.02 | 726.71 | 744.17 |
| 11 | Isobutyl butyrate | 940.26 | 956.57 | 935.85 | 951.45 | 942.12 | 962.02 | 940.19 | 951.73 |
| 12 | 2-Methyl-3-pentanone 7 | 733.02 | 752.40 | 730.85 | 750.29 | 642.27 | 675.20 | 729.12 | 753.16 |
| 13 | 2,6-Dimethyl-4-heptanone | 954.66 | 970.95 | 987.70 | 999.46 | 983.67 | 991.68 | 990.96 | 981.93 |
| 14 | 2,3-Dimethyl-2-butanol | 715.26 | 729.44 | 714.70 | 737.64 | 635.23 | 672.76 | 706.62 | 734.66 |
| 15 | 3-Hexanol | 780.36 | 795.07 | 776.55 | 797.04 | 711.05 | 732.12 | 769.10 | 803.06 |
| 16 | 3,5-Dimethyl-3-hexanol | 883.13 | 896.48 | 881.05 | 900.30 | 890.45 | 910.45 | 880.08 | 893.57 |
| 17 | 3-Octanol | 981.97 | 996.71 | 962.43 | 978.90 | 962.95 | 981.74 | 968.49 | 985.66 |

**Table 1:** Data set and corresponding experimental (exp.) and predicted (cal.) values of RI.

for the first position, where only functions are allowed). However, in the tail of a gene only terminal is allowed.

Mutation, Inversion, Transposition and Recombination

Mutation: Mutations can occur anywhere in the chromosome. Simple mutation just replaces symbols in genes with replacement symbols. However, the structural organization of chromosomes must remain intact. Symbols in the heads of genes can be replaced by functions or terminals (variables and constants). Symbols in the tail sections can be replaced only by terminals. Randomly change symbols in a chromosome. Symbols in the tail of a gene may not operate on any arguments. Typically two-point mutation per chromosome is used. It is worth noticing that in GEP there are no constraints neither in the kind of mutation nor the number of mutations in a chromosome: in all cases the newly created individuals are syntactically correct programs.

Inversion: Inversion reverses the order of symbols in a section of a gene. A portion of a chromosome is chosen to be inserted in the head of a gene. The tail of the gene is unchanged. Thus symbols are removed from the end of the head to make room for the inserted string. Typically a probability of 0.1 of insertion is used.

Transposition: Transposition selects a group of symbols and moves the symbols to a different position within the same gene. Gene transposition moves entire genes around in the chromosome. One gene in a chromosome is randomly chosen to be the first gene. All other genes in the chromosome are shifted downwards in the chromosome to make place for the first gene.

An IS element is a variable-size sequence of elements extracted from a random starting point within the genome (even if the genome was composed of several chromosomes). Another position within the genome is chosen as the insertion point.

This target site must be within the head part of a gene and cannot be the first element (gene root). The IS element is sequentially inserted in the target site, shifting all elements from this point onwards and a sequence with the same number of elements is deleted from the end of the head, so that the structural organization is maintained. This operator simulates the transposition found in the evolution of biological genomes. RIS is similar to the IS transposition, except that the insertion sequence must have a function as the first element and the target point must be also the first element of a gene (root).

The transposable elements of GEP are fragments of the genome that can be activated and jump to another place in the chromosome: (1) Short fragments with a function or terminal in the first position that transpose to the head of genes, except to the root (insertion sequence elements or IS elements); (2) Short fragments with a function in the first position that transpose to the root of genes (root IS elements or RIS elements); (3) Entire genes that transpose to the beginning of chromosomes.

Recombination: During recombination, two chromosomes are randomly selected, and genetic material is exchanged between them to produce two new chromosomes.

The cross over operation this can be one point (the chromosomes are split in two and corresponding sections are swapped), two point (chromosomes are split in three and the middle portion is swapped) or gene (one entire gene is swapped between chromosomes) recombination. Typically the sum of the probabilities of recombination is 0.7.

In GEP there are three kinds of recombination: one-point, two-point, and gene recombination. (1) One-point: During one-point recombination, the chromosomes crossover a randomly chosen point to form two daughter chromosomes; (2) Two-point: In two-point recombination the chromosomes are paired and the two points of recombination are randomly chosen. The material between the recombination points is afterwards exchanged between the two chromosomes, forming two new daughter chromosomes; (3) Gene recombination: recombines entire genes. This operator randomly chooses genes in the same position in two parent chromosomes to form two new off springs. In gene recombination an entire gene is exchanged during crossover. The exchanged genes are randomly chosen and occupy the same position in the parent chromosomes. It is worth noting that this operator is unable to create new genes: the individuals created are different arrangements of existing genes.

**Fitness function:** A fitness function is the most important part of any EA application. Fitness function given with above equations allows for fulfilling all of the set conditions. In GEP, fitness is based on how well an individual model the data. If the target variable has continuous values, the fitness can be based on the difference between predicted values and actual values. Evolution stops when the fitness of the best individual in the population reaches some limit that is specified for the analysis or when a specified number of generations have been created or a maximum execution time limit is reached.

All of the fitness functions produce fitness scores in the range 0.0 to 1.0 with 1.0 being ideal fitness – that is, the individual exactly fits the data. If a function is unviable – for example, it takes the square root of a negative number or divides by zero – then its fitness score is 0.0.

**GEP evolution process:** The GEP evolution begins with the random generation of linear fixed-length chromosomes for individuals of the initial population. The chromosomes are translated into ETs and subsequently into mathematical expressions, and the fitness of each individual is evaluated based on a pre-defined fitness function. The individuals are then selected by fitness to reproduce with modification. The individuals of this new generation are, in their run, subject to the same developmental process. The selection and reproduction is accomplished by roulette-wheel sampling with elitism, which guarantees the survival and cloning of the best individual to the next generation. Variation in the population is introduced by applying one or more genetic operators to selected chromosomes, including crossover, mutation and insertion.

## Models

**GEP model:** The GEP program was coded by the combination of MATLAB and VC++. The MATLAB software has the advantage of computing matrix conveniently and programming efficiently, but its operating efficiency is relatively low. So VC++ was combined for its powerful function and the characteristics of higher operating efficiency with MATLAB. In this paper, MATLAB engine was used to achieve the combination with VC++ programming. There are two steps: (1) Add MATLAB engine library header files and library functions of the path. (2) Add libmx.lib libeng.lib libmex.lib to complete the import of the corresponding MATLAB engine static link library.

From the data in Table 1, GEP method was used that 6 topological index as input, output for its retention index. During the run, parameter values were needed to adjust constantly in order to achieve the optimal results. The set of optimal parameter values were listed in Table 2 and the predicting results of test set on OV-1 and SE-54 were listed in Figures 1 and 2. It can be seen from the figures that the predictive values of gas chromatography retention index of oxygen-organic compounds

| Parameters | Values |
|---|---|
| Generation | 2000 |
| Population Size | 100 |
| Function Set | "+" "-" "*" "/" "sin" "cos" "sqrt" "exp" "ln" |
| Head Size | 8 |
| Number Of Genes | 3 |
| Linking Function | + |
| Mutation Rate | 0.044 |
| 1-Point recombination rate | 0.3 |
| Gene recombination | 0.3 |
| Gene | 0.1 |
| IS transposition rate | 0.1 |
| RIS transposition rate | 0.1 |
| Gene transposition rate | 0.1 |
| Selection range | 100 |

**Table 2:** Parameters of GEP models.

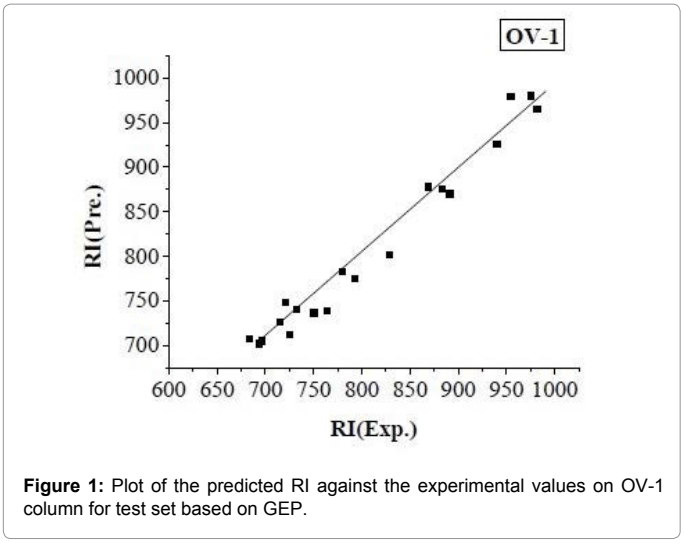



**Figure 1:** Plot of the predicted RI against the experimental values on OV-1 column for test set based on GEP.

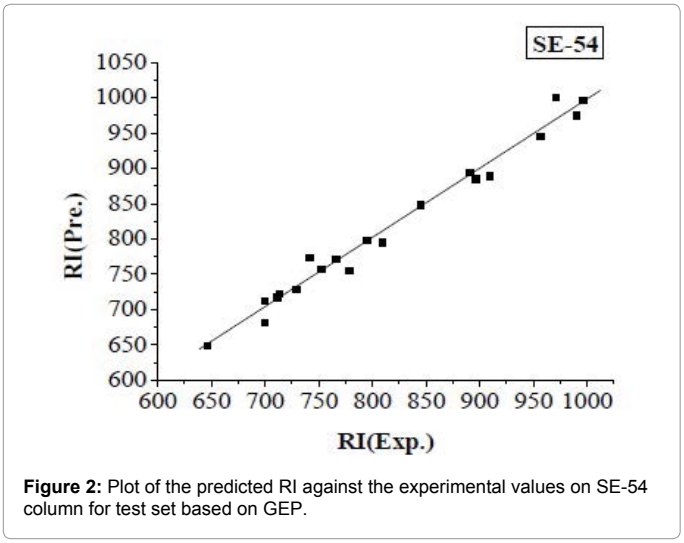



**Figure 2:** Plot of the predicted RI against the experimental values on SE-54 column for test set based on GEP.

were in good agreement with the experimental data.

**ANN model:** Non-linear statistical treatment of QSRR data is expected to provide models with better predictive quality as compared with related MLR models. In this perspective, functioning and applications of ANN have been adequately described elsewhere [7-10]. Extensive use of ANN, which has inherent ability to incorporate nonlinear and cross-product terms into the model and does not require prior knowledge of the mathematical function as well, largely rests on its flexibility and less sensitivity to collinearity among variables. The theory behind ANN and their use in chromatography have been reported elsewhere [11-13].

Multi-layer feed forward networks, with good self-learing ability and adaptability is widely used in the field of QSRR modeling [14]. Commonly, they consist of three layers: one input layer formed by a number of neurons that equal to the number of descriptors, one out neuron (providing the model response) and a number of hidden neurons fully connected to both input and out neurons. Among the available learning algorithms, back-propagation of errors is one of the most widely used [8,15].

Usually, there are four steps involved in ANN modeling: (1) assembling the training data of input (independent variables) and output (dependent variables), (2) deciding the network architecture, (3) training the network, and (4) simulating the network response to new inputs. The training process is simply an optimization process which aims at finding the set of weight and biases associated with each layer that will minimize the error objective function related to the deviations of the network predictions from the true response output data of the training set.

Before data set was used for the training of ANN, it was normalized separately. Its minimum value was set to zero and maximum to one. The proper number of nodes in the hidden layer was determined by training the network with different number of nodes in the hidden layer. The root-mean-square error (RMSE) value measures how good the outputs are in comparison with the target values. In this paper, following a troubleshooting study to investigate the effects of the number of hidden layers and the number of neurons involved in these hidden layers, a 2-3-1 network, with tansig-logsig transfer functions, was found to be the most optimum in terms of the root mean squared errors (RMSE) obtained.

ANN with basic back-propagation of errors learning algorithm was used in this study to predict oxygen-containing retention index. A three-layer network with a sigmoid transfer function was designed for ANN. The ANN program was coded in MATLAB 7 for windows [15].

**The MLR:** For regression analysis, data set was randomly divided into two groups: training and test sets. The training set, composed of 74 molecules, was used for the model generation. The test set, composed of 17 molecules, was used to evaluate the generated model. The program used for MLR analysis was compiled in Statistical Product and Service Solutions (SPSS version 19.0 IBM) software. In MLR analysis, in order to minimize the information overlap in descriptors and to reduce the number of descriptors required in regression equation, the concept of non-redundant descriptors was used in this study. The best equation was selected on the basis of the highest multiple correlation coefficients (R) and the lowest root mean squared error (RMS). The linear equation between these descriptors and the retention parameters of fluid catalytic cracking (FCC) gasoline was:

$$y = b_0 + b_1 x_1 + b_2 x_2 + .... + b_p x_p \quad (1)$$

Where $b_0$ is the intercept and $b_j$ is the regression coefficient for descriptor j. The statistical results obtained by using the two molecular descriptors based on MLR are listed in Table 3 and plotted against the experimental values in Figures 3 and 4.

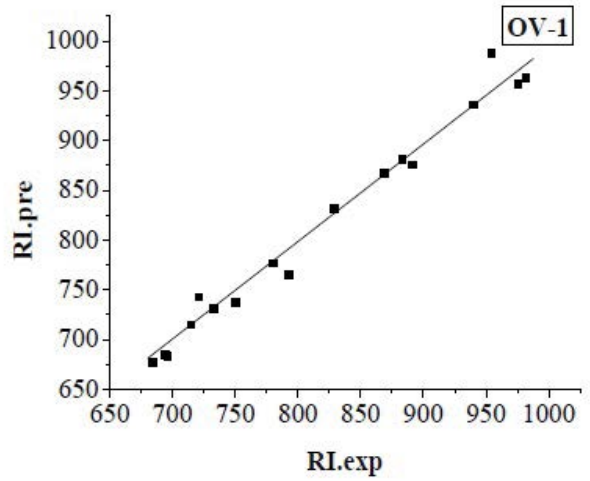



**Figure 3:** Plot of the predicted RI against the experimental values on OV-1 column for test set based on MLR.

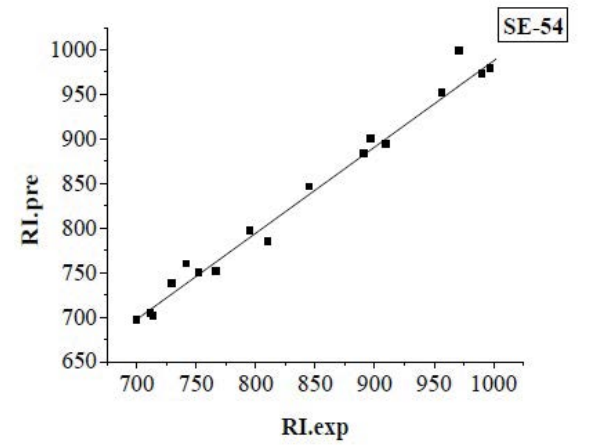



**Figure 4:** Plot of the predicted RI against the experimental values on SE-54 column for test set based on MLR.

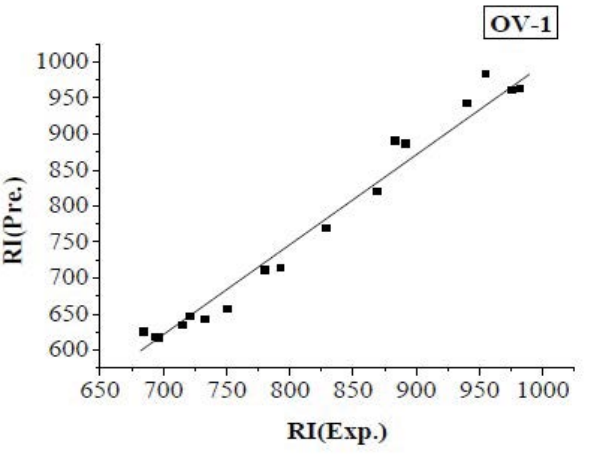



**Figure 5:** Plot of the predicted RI against the experimental values on OV-1 column for test set based on ANN.

It is common to consider four statistical parameters for regression equation. These parameters are the number of descriptors, correlation coefficient (R) for training and test sets, root mean squared error (RMS) for training and test sets, and F statistic. A reliable MLR model is one that has high R and F values, low RMS and least number of descriptors. In addition to these, the model should have a high predictive ability. Consequently, among different models, the best model was chosen, whose specifications are presented in Table 3. Here the corresponding descriptors used in MLR were applied as inputs for ANN in order to compare the performance of the two models.

## Results

The main aim of the present work was developing a QSRR model to predict the retention parameter (RI) of oxygen-containing compounds appeared in Table 1. A linear model of MLR was developed, whose specifications are given in Table 3. All statistic tests were performed at a significance level of 5%. MLR model performance was measured by three metrics: (1) R, which gives the fraction explained variance for the analyzed set, was used to measure the model's fit performance. (2) Root Mean Squared error (RMS), which can give the bias in the prediction, was used to evaluate the model's predictive precision: the lower the RMS, the better the prediction precision. It can be calculated as below:

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(d_i - o_i)^2}{n}} \quad (3)$$

where $d_i$ is the target value, $o_i$ is the experimental value and n is the number of compounds in analyzed set. (3) The variance ratio of calculated and observed activities F.

After the linear model was gained, non-linear characteristics of the retention parameter were also performed using ANN. Here a feed-forward neural network with basic error back-propagation algorithm was constructed to model the nonlinear QSRR models. Therefore, a 2-3-1 BP-ANN, with tansig-logsig transfer functions, was developed. Figure 5 demonstrates the plot of the ANN predicted versus the experimental values of the RI for the data set. A correlation coefficient of this plot indicates the reliability of the model. As can been seen in Table 4, the correlation coefficient R on OV-1 and SE-54 for the ANN models are larger than that of MLR models respectively, which indicates that the ANN models are slightly improved to MLR models. The residuals of calculated values of RI by ANN are plotted against the experimental values in Figure 6. The propagation of the residuals on both sides of zero line indicates that no symmetric error exists in the development of ANN model.

| Descriptors | OV-1 coefficient | Std. Error | SE-54 coefficient | Std. Error | OV-1 test values | SE-54 test value |
|---|---|---|---|---|---|---|
| Constant | 1951.898 | 450.621 | 2028.14 | 345.24 | 4.332 | 5.875 |
| OEI | 48.089 | 4.351 | 50.562 | 3.333 | 11.053 | 15.169 |
| SX1CH | -43.038 | 151.408 | 3.36 | 116 | -0.284 | 0.029 |
| N2/3 | 3.331 | 51.507 | 26.672 | 39.462 | 0.065 | 0.676 |
| χeq × PEI | -459.326 | 162.605 | -504.236 | 124.579 | -2.825 | -4.048 |
| χeq | -173.948 | 12.183 | -159.685 | 9.334 | -14.278 | -17.108 |
| MPEIm × IMPEIm | 5.589 | 1.039 | 5.694 | 0.796 | 5.382 | 7.156 |

**Table 3:** Model parameters value and coefficients for MLR model.

| OV-1 | Test sets | SE-54 | Test sets |
|---|---|---|---|
| **Method** | **R** | **Method** | **R** |
| MLR | 0.9911 | MLR | 0.9917 |
| ANN | 0.9891 | ANN | 0.9892 |
| GEP | 0.9909 | GEP | 0.9955 |

**Table 4:** Result of correlation coefficient (R) with MLR, ANN and GEP for the test set.

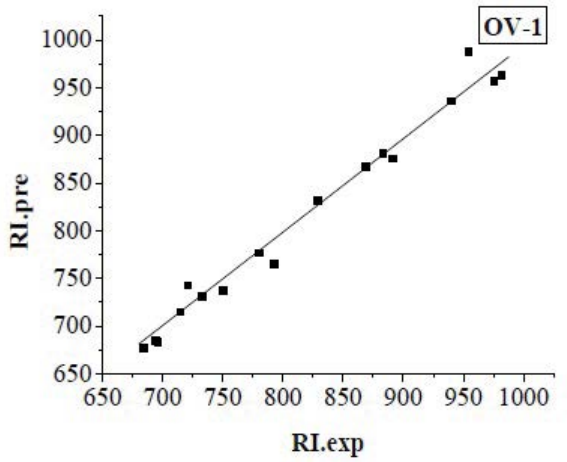



**Figure 6:** Plot of the predicted RI against the experimental values on SE-54 column for test set based on ANN.

## Acknowledgements

Correspondence author is grateful to the financial support from National Natural Science Foundation of China, Lanzhou University of China and Liaoning Province Education Department of China, which made this work possible.

### References

1. Lu C, Abrahan F, Adamowiciz L (2007) QSRR study for gas and liquid chromatographic retention indices of polyhalogenated biphenyls using two 2D decriptors. Chromatographia 66: 717-724.
2. Daghir-Wojtkowiak E, Studzi´nska S, Buszewski B, Kaliszanc R, JanMarkuszewski M (2014) Quantitative structure–retention relationships of ionic liquid cations in characterization of stationary phases for HPLC. Royal Society of Chemistry 6: 1189-1196.
3. Ghasemi J, Saaidpour S, Brown SD (2007) QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. J Mol Struct Theochem 805: 27-32.
4. Liu F, Liang Y, Cao C, Zhou N (2007) Theoretical prediction of the Kovat's retention index for oxygen-containing organic compounds using novel topological indices. Anal Chim Acta 594: 279-289.
5. Ferreira C (2006) Gene Expression Programming: Mathematical Modelling by an Artificial Intelligence, Angra do Heroismo, Portugal.
6. Zhang XM, Lu PC (1996) Unified equation between Kováts indices on different stationary phases for select types of compounds. J Chromatogr A 731: 187-199.
7. Noorizaden H, Farmany A (2010) QSRR models to predict rentention indices of cyclic compounds of essential oils. J Chemometrics Laboratory 72: 563-569.
8. Gorynski K, Bojko B, Nowaczyk A, Bucinski A, Pawliszyn J, et al. (2013) Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds. Analytical Chimica Acta 797: 13-19.
9. Mior R, Morés S, Welz B, Carasek E, de Andrade JB (2013) Determination of sulfur in coal using direct solid sampling and high-resolution continuum source molecular absorption spectrometry of the CS molecule in a graphite furnace. Talanta 106: 368-374.
10. Gupta VK, Khani H, Ahmadi-Roudi B, Mirakhorli S, Fereyduni E, et al. (2011) Prediction of capillary gas chromatographic retention times of fatty acid methyl esters in human blood using MLR, PLS and back-propagation artificial neural networks. Talanta 83: 1014-1022.
11. D'Archivio AA, Incani A, Ruggieri F (2011) Retention modeling of polychlorinated biphenyls in comprehensive two-dimensional gas chromatography. Anal Bioanal Chem 399: 903-913.
12. Fatemi MH (2002) Simultaneous modeling of the Kovats retention indices on OV-1 and SE-54 stationary phases using artificial neural networks. J Chromatogr A 955: 273-280.
13. Zhang X, Zhang X, Li Q, Sun Z, Song L, et al. (2014) Support Vector Machine Applied to Studyon Quantitative Structure–Retention Relationships of Polybrominated Diphenyl Ether Congeners. Chromatographia 77: 1387-1398.
14. Xiaotong Z, Xingming L (2004) GBP network application in the modified paraffin wax quality prediction. Journal of Petrochemical Universities 21: 1-5.
15. MATLAB, The Language of Technical Computing.