Int. J. Mol. Sci. **2009**, *10*, 1978-1998; doi:10.3390/ijms10051978

OPEN ACCESS

International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Review

Current Mathematical Methods Used in QSAR/QSPR Studies

Peixun Liu [#] and Wei Long [#]

Institute of Radiation Medicine, Peking Union Medical College, Chinese Academy of Medical Sciences, Tianjin 300192, P.R. China; E-Mails: longwaylong@gmail.com (W.L.); pharm8888@yahoo.com.cn (P.-X.L.); Tel. +86-22-85683042; Fax: +86-22-85682394

[#] These authors contributed equally to this work

Received: 19 March 2009 / Accepted: 28 April 2009 / Published: 29 April 2009

Abstract: This paper gives an overview of the mathematical methods currently used in quantitative structure-activity/property relationship (QASR/QSPR) studies. Recently, the mathematical methods applied to the regression of QASR/QSPR models are developing very fast, and new methods, such as Gene Expression Programming (GEP), Project Pursuit Regression (PPR) and Local Lazy Regression (LLR) have appeared on the QASR/QSPR stage. At the same time, the earlier methods, including Multiple Linear Regression (MLR), Partial Least Squares (PLS), Neural Networks (NN), Support Vector Machine (SVM) and so on, are being upgraded to improve their performance in QASR/QSPR studies. These new and upgraded methods and algorithms are described in detail, and their advantages and disadvantages are evaluated and discussed, to show their application potential in QASR/QSPR studies in the future.

Keywords: QSAR; QSPR; Mathematical methods; Regression; Algorithm.

1. Introduction

As a common and successful research approach, quantitative structure activity/property relationship (QASR/QSPR) studies are applied extensively to chemometrics, pharmacodynamics, pharmacokinetics, toxicology and so on. Recently, the mathematical methods used as regression tools in QSAR/QSPR analysis have been developing quickly. Thus, not only are the previous methods, such as Multiple Linear Regression (MLR), Partial Least Squares (PLS), Neural Networks (NN), Support Vector Machine (SVM), being upgraded by improving the kernel algorithms or by combining them with other methods, but also some new methods, including Gene Expression Programming (GEP), Project Pursuit Regression (PPR) and Local Lazy Regression (LLR), are being mentioned in the current reported QSAR/QSPR studies. In light of this, this paper is divided as follows: first, the improved methods based on the earlier ones are described, then the new methods are presented. Finally, all these methods are jointly commented and their advantages and disadvantages discussed to show their application potential in future QASR/QSPR studies.

2. Multiple Linear Regression (MLR)

MLR is one of the earliest methods used for constructing QSAR/QSPR models, but it is still one of the most commonly used ones to date. The advantage of MLR is its simple form and easily interpretable mathematical expression. Although utilized to great effect, MLR is vulnerable to descriptors which are correlated to one another, making it incapable of deciding which correlated sets may be more significant to the model. Some new methodologies based on MLR have been developed and reported in recent papers aimed at improving this technique. These methods include Best Multiple Linear Regression (BMLR), Heuristic Method (HM), Genetic Algorithm based Multiple Linear Regression (GA-MLR), Stepwise MLR, Factor Analysis MLR and so on. The three most important and commonly used of these methods are described in detail below.

2.1. Best Multiple Linear Regression (BMLR)

BMLR implements the following strategy to search for the multi-parameter regression with the maximum predicting ability. All orthogonal pairs of descriptors i and j (with R2ij < R2min, default value R2ij < 0.1) are found in a given data set. The property analyzed is treated by using the two-parameter regression with the pairs of descriptors, obtained in the first step. The Nc (default value Nc = 400) pairs with highest regression correlation coefficients are chosen for performing the higher-order regression treatments. For each descriptor pair, obtained in the previous step, a non-collinear descriptor scale, k (with R2ik < R2nc and R2kj < R2nc, default value R2 < 0.6) is added, and the respective three-parameter regression treatment is performed. If the Fisher criterion at a given probability level, F, is smaller than that for the best two-parameter correlation, the latter is chosen as the final result. Otherwise, the Nc (default value Nc = 400) descriptor triples with highest regression correlation coefficients are chosen for the next step. For each descriptor set, chosen in the previous step, an additional non-collinear descriptor scale is added, and the respective *(n + 1)*-parameter regression treatment is performed. If the Fisher criterion at the given probability level, *F*, is smaller than for the best two-parameter correlation, the latter is chosen as the final result. Otherwise, the *Nc* (default value *Nc* = 400) sets descriptor sets with highest regression correlation coefficients are chosen, and this step repeated with $n = n + 1$ [1].

As an improved method based on MLR, BMLR is instrumental for variable selection and QSAR/QSPR modeling [2-8]. Like MLR, BMLR is noted for its simple and interpretable mathematical expression. Moreover, overcoming the shortcomings of MLR, BMLR works well when the number of compounds in the training set doesn't exceed the number of molecular descriptors by at least a factor of five. However, BMLR will derive an unsatisfactory result when the structure-activity relationship is non-linear in nature. When too many descriptors are involved in a calculation, the modeling process will be time consuming. To speed up the calculations, it is advisable reject descriptors with insignificant variance within the dataset. This will significantly decrease the probability of including unrelated descriptors by chance. In addition, BMLR is unable to build a one-parameter model. BMLR is commercially available in the software packages CODESSA [9] or CODESSA PRO [10].

2.2. Heuristic Method (HM)

HM, an advanced algorithm based on MLR, is popular for building linear QSAR/QSPR equations because of its convenience and high calculation speed. The advantage of HM is totally based on its unique strategy of selecting variables. The details of selecting descriptors are as follows: first of all, all descriptors are checked to ensure that values of each descriptor are available for each structure. Descriptors for which values are not available for every structure in the data are discarded. Descriptors having a constant value for all structures in the data set are also discarded. Thereafter all possible one-parameter regression models are tested and the insignificant descriptors are removed. As a next step, the program calculates the pair correlation matrix of descriptors and further reduces the descriptor pool by eliminating highly correlated descriptors. The details of validating intercorrelation are: (a) all quasi-orthogonal pairs of structural descriptors are selected from the initial set. Two descriptors are considered orthogonal if their intercorrelation coefficient r_{ij} is lower than 0.1; (b) the pairs of orthogonal descriptors are used to compute the biparametric regression equations; (c) to a multi-linear regression (MLR) model containing *n* descriptors, a new descriptor is added to generate a model with $n + 1$ descriptors if the new descriptor is not significantly correlated with the previous *n* descriptors; step (c) is repeated until MLR models with a prescribed number of descriptors are obtained. The goodness of the correlation is tested by the square of coefficient regression (R^2), square of cross-validate coefficient regression (q^2), the F-test (F), and the standard deviation (S) [1].

HM is commonly used in linear QSAR and QSPR studies, and also as an excellent tool for descriptor selection before a linear or nonlinear model is built [11-34]. The advantages of HM are the high speed and the absence of software restrictions on the size of the data set. HM can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. HM usually produces correlations 2 – 5 times faster than other methods with comparable quality. Additionally, the maximum number of parameters in the resulting model can be fixed in accordance with the situation so as to save time. As a method inherited from MLR, HM is also limited in linear models.

2.3. Genetic Algorithm based Multiple Linear Regression (GA-MLR)

Combining Genetic Algorithm (GA) with MLR, a new method called GA-MLR is becoming popular in currently reported QSAR and QSPR studies [35-55]. In this method, GA is performed to search the feature space and select the major descriptors relevant to the activities or properties of the compounds. This method can deal with z large search space efficiently and has less chance to become a local optimal solution than the other algorithms. We give a brief summary of the main procedure of GA herein. The first step of GA is to generate a set of solutions (chromosomes) randomly, which is called an initial population. Then, a fitness function is deduced from the gene composition of a chromosome. The Friedman LOF function is commonly used as the fitness function, which was defined as follows:

$$\mathrm{LOF} = \{\mathrm{SSE}/(1-(c+dp/n))\}^2 \quad (1)$$

where SSE is the sum of squares of errors, *c* is the number of the basis function (other than the constant term), *d* is the smoothness factor, *p* is the number of features in the model, and *n* is the number of data points from which the model is built. Unlike the R^2 error, the LOF measure cannot always be reduced by adding more terms to the regression model. By limiting the tendency to simply add more terms, the LOF measure resists over-fitting of a model. Then, crossover and mutation operations are performed to generate new individuals. In the subsequent selection stage, the fittest individuals evolve to the next generation. These steps of evolution continue until the stopping conditions are satisfied. After that, the MLR is employed to correlate the descriptors selected by GA and the values of activities or properties.

GA, a well-estimated method for parameter selection, is embedded in GA-MLR method so as to overcome the shortage of MLR in variable selection. Like the MLR method, the regression tool in GA-MLR, is a simple and classical regression method, which can provide explicit equations. The two parts have a complementation for each other to make GA-MLR a promising method in QSAR/QSPR research.

3. Partial Least Squares (PLS)

The basic concept of PLS regression was originally developed by Wold [56,57]. As a popular and pragmatic methodology, PLS is used extensively in various fields. In the field of QSAR/QSPR, PLS is famous for its application to CoMFA and CoMSIA. Recently, PLS has evolved by combination with other mathematical methods to give better performance in QSAR/QSPR analyses. These evolved PLS', such as Genetic Partial Least Squares (G/PLS), Factor Analysis Partial Least Squares (FA-PLS) and Orthogonal Signal Correction Partial Least Squares (OSC-PLS), are briefly introduced in the following sections.

3.1. Genetic Partial Least Squares (G/PLS)

G/PLS is derived from two QSAR calculation methods Genetic Function Approximation (GFA) [58,59] and PLS. The G/PLS algorithm uses GFA to select appropriate basis functions to be used in a model of the data and PLS regression is used as the fitting technique to weigh the basis functions' relative contributions in the final model. Application of G/PLS thus allows the construction of larger QSAR equations while still avoiding over-fitting and eliminating most variables. As the regression method used in Molecular Field Analysis (MFA), a well-known 3D-QSAR analysis tool, G/PLS is commonly used. The recent literatures related to G/PLS are mainly listed as [60-70].

3.2. Factor Analysis Partial Least Squares (FA-PLS)

This is the combination of Factor Analysis (FA) and PLS, where FA is used for initial selection of descriptors, after which PLS is performed. FA is a tool to find out the relationships among variables. It reduces variables into few latent factors from which important variables are selected for PLS regression. Most of the time, a leave-one-out method is used as a tool for selection of optimum number of components for PLS. We can find examples of FA-PLS used in QSAR analysis in [68,71-74].

3.3. Orthogonal Signal Correction Partial Least Squares (OSC-PLS)

Orthogonal signal correction (OSC) was introduced by Wold *et al.* [75] to remove systematic variation from the response matrix X that is unrelated, or orthogonal, to the property matrix Y. Therefore, one can be certain that important information regarding the analyte is retained. Since then, various OSC algorithms have been published in an attempt to reduce model complexity by removing orthogonal components from the signal. *In abstracto*, a preprocessing with OSC will help traditional PLS to obtain a more precise model, as proven in many studies of spectral analysis [76-88]. To date, unfortunately, there are only a few reports in which OSC-PLS is applied to QSAR/QSPR studies [89-91], but more QSAR or QSPR research involving application of the OSC-PLS method are expected in the future.

4. Neural Networks (NN)

As an alternative to the fitting of data to an equation and reporting the coefficients derived therefrom, neural networks are designed to process input information and generate hidden models of the relationships. One advantage of neural networks is that they are naturally capable of modeling nonlinear systems. Disadvantages include a tendency to overfit the data, and a significant level of difficulty in ascertaining which descriptors are most significant in the resulting model. In the recent QSAR/QSPR studies, RBFNN and GRNN are the most frequently used ones among NN.

4.1. Radial Basis Function Neural Network (RBFNN)

The RBFNN consists of three layers: an input layer, a hidden layer and an output layer. The input layer does not process the information; it only distributes the input vectors to the hidden layer. Each neuron on the hidden layer employs a radial basis function as a nonlinear transfer function to operate on the input data. In general, there are several radial basis functions (RBF): linear, cubic, thin plate spline, Gaussian, multi-quadratic and inverse multi-quadratic. The most often used RBF is a Gaussian function that is characterized by a center (c_j) and width (r_j). Due to the limited length of writing, only Gaussian RBF is introduced in this paper. The nonlinear transformation with RBF in the hidden layer is given as follows:

$$h_j(x)=\exp\left(\frac{-\left\|x-c_j\right\|^2}{r_j^2}\right) \tag{2}$$

where, h_j is the notation for the output of the jth RBF unit, c_j and r_j are the center and width of the jth RBF, respectively. The operation of the output layer is linear, which is given as below:

$$y_k(x)=\sum w_{kj}h_j(x)+b_k \tag{3}$$

where, y_k is the kth output unit for the input vector x, w_{kj} is the weight connection between the kth output unit and the jth hidden layer unit, and b_k is the bias. The training procedure when using RBF involves selecting centers, width and weights. In this paper, the forward subset selection routine was used to select the centers from training set samples. The adjustment of the connection weight between the hidden layer and the output layer was performed using a least squares solution after the selection of centers and widths of radial basis functions. Compared with the Back Propagation Neural Network (BPNN), RBFNN has the advantage of short training time and is guaranteed to reach the global minimum of error surface during training process. The optimization of its topology and learning parameters are easy to implement. Applications of RBFNN in QSAR/QSPR studies can be found in [22,24,27,32,51,92-102].

4.2. General Regression Neural Network (GRNN)

GRNN, one of the so-called Bayesian networks, is a type of neural network using kernel-based approximation to perform regression. It was introduced by Specht in 1991 [103]. GRNN is a nonparametric estimator that calculates a weighted average of the target values of training patterns by the probability density function using Parzen's nonparametric estimator. For GRNN, the predicted value is the most probable value $E(y|x)$:

$$E(y\,|x)\overset{\square}{=}\hat{y}(x)=\frac{\int_{-\infty}^{+\infty} yf(x,y)\mathrm{d}y}{\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y} \tag{4}$$

where $f(x, y)$ is the probability density function. This can be estimated from the training set by using the Parzen's nonparametric estimator [47]:

$$f(x,y)=\frac{1}{n\sigma}\sum_{i=1}^{n} w\frac{(x-x_i)}{\sigma} \tag{5}$$

where n is the sample size, s is a scaling parameter that defines the width of the bell curve that surrounds each sample point, $W(d)$is a weighting function that has its largest value at d=0, and (x-x_i) is the distance between the unknown sample and a data point. The Gaussian function is frequently used as the weighting function because it is well behaved, easily calculated, and satisfies the conditions required by Parzen's estimator. Substituting Parzen's nonparametric estimator for $f(x, y)$ and performing the integrations leads to the fundamental equation of GRNN:

$$\hat{y}(x)=\frac{\sum_{i=1}^{n} y_i\exp(-D(x,x_i))}{\sum_{i=1}^{n}\exp(-D(x,x_i))} \tag{6}$$

where:

$$D(x,x_i)=\sum_{j=1}^{p}\left(\frac{x_j-x_{ij}}{\sigma_j}\right)^2 \tag{7}$$

GRNN consists of four layers: input, hidden, summation, and output layers. The greatest advantage of GRNN is the training speed. Meanwhile, it is relatively insensitive to outliers (wild points). Training a GRNN actually consists mostly of copying training cases into the network, and so is as close to instantaneous as can be expected. The greatest disadvantage is network size: a GRNN network actually contains the entire set of training cases, and is therefore space-consuming, requiring more memory space to store the model. Relative literatures are [46,104-109].

5. Support Vector Machine (SVM)

SVM, developed by Vapnik [110,111] as a novel type of machine learning method, is gaining popularity due to its many attractive features and promising empirical performance. Originally, SVM was developed for pattern recognition problems. After that, SVM it was applied to regression by the introduction of an alternative loss function and results appear to be very encouraging [112]. As a developing method, new types of SVM are coming in on the stage of QSAR/QSPR, such as Least Square Support Vector Machine (LS-SVM), Grid Search Support Vector Machine (GS-SVM), Potential Support Vector Machine (P-SVM) and Genetic Algorithms Support Vector Machine (GA-SVM). Here, we only choose LS-SVM, the most commonly used one, as an example to provide a description.

5.1. Least Square Support Vector Machine (LS-SVM)

The LS-SVM, which was a modified SVM algorithm, was proposed by Suykens *et al.* [113] and used to build the nonlinear model. Here, we only briefly describe the main idea of LS-SVM for function estimation. In principle, LS-SVM always fits a linear relation (y = wx + b) between the regressors (x) and the dependent variable (y). The best relation can be obtained by minimizing the cost function (Q) containing a penalized regression error term:

$$Q_{\text{LS-SVM}}=\frac{1}{2}w^Tw+\gamma\sum_{k=1}^{N}e_k{}^2 \tag{8}$$

subject to:

$$y_k=w^T\varphi(x_k)+b+e_k\,,\ k=1,\ \dots\ ,\ N \tag{9}$$

where φ: $R^n \rightarrow R^m$ is the feature map mapping the input space to a usually high-dimensional feature space, γ is the relative weight of the error term, and e_k is error variables taking noisy data into accurate and avoiding poor generalization. LS-SVM considers this optimization problem to be a constrained optimization problem and uses a language function to solve it. By solving the Lagrange style of equation (8), the weight coefficients (w) can be written as:

$$w=\sum_{k=1}^{N}\alpha_k x_k^T x+b \text{ with } \alpha_k=2\gamma e_k \tag{10}$$

By substituting it into the original regression line (y □ wx + b), the following result can be obtained:

$$y = \sum_{k=1}^{N} \alpha_k x_k^T x + b \tag{11}$$

It can be seen that the Lagrange multipliers can be defined as:

$$\alpha_k = (x_k^T x + (2\gamma)^{-1}(y_k - b))$$

Finding these Lagrange multipliers is very simple, as opposed to with the SVM approach, in which a more difficult relation has to be solved to obtain these values. In addition, it easily allows for a nonlinear regression as an extension of the linear approach by introducing the kernel function. This leads to the following nonlinear regression function:

$$f(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \tag{12}$$

where $K(x, x_k)$ is the kernel function. The value is equal to the inner product of two vectors x and x_k in the feature space $\Phi(x)$ and $\Phi(x_k)$; that is, $K(x, x_k)$) = $\Phi(x)^T\Phi(x_k)$. The choices of a kernel and its specific parameters together with γ have to be tuned by the user. The radial basis function (RBF) kernel $K(x, x_k) = \exp(-\|x_k - x\|^2/\sigma^2)$ is commonly used, and then leave-one-out (LOO) cross-validation is employed to tune the optimized values of the two parameters γ and σ. LS-SVM is a simplification of traditional support vector machine (SVM). It encompasses similar advantages with SVM and its own additional advantages. It only requires solving a set of linear equations (linear programming), which is much easier and computationally simpler than nonlinear equations (quadratic programming) employed by traditional SVM. Therefore, LS-SVM is faster than traditional SVM in treating the same work. The related literature is presented in [37,114-118].

6. Gene Expression Programming (GEP)

Gene expression programming was invented by Ferreira in 1999 and was developed from genetic algorithms and genetic programming (GP). GEP uses the same kind of diagram representation of GP, but the entities evolved by GEP (expression trees) are the expression of a genome. GEP is more simple than cellular gene progression. It mainly includes two sides: the chromosomes and the expression trees (ETs). The process of information of gene code and translation is very simple, such as a one-to-one relationship between the symbols of the chromosome and the functions or terminals they represent. The rules of GEP determine the spatial organization of the functions and terminals in the ETs and the type of interaction between sub-ETs. Therefore, the language of the genes and the ETs represents the language of GEP.

6.1. The GEP chromosomes, expression trees (ETs), and the mapping mechanism

Each chromosome in GEP is a character string of fixed length, which can be composed of gene from the function set or the terminal set. Using the elements {+, -,*, /, *Q*} as the function set and {*a, b, c, d*} as the terminal set, the following is an example of GEP chromosome of length eight:

$$\left\{ \begin{array}{c} 01234567 \\ Q{*}{+}{-}abcd \end{array} \right\}$$

where Q denotes the square-root function; and *a, b, c, d* are variable (or attribute) names. The above is referred to as Karva notation or *K*-expression. A *K*-expression can be mapped into an ET following a width-first procedure. A branch of the ET stops growing when the last node in this branch is a terminal. The conversion of an ET into a *K*-expression is also very straightforward and can be accomplished by recording the nodes from left to right in each layer of the ET in a top-down fashion to form the string:

$$y = \sqrt{(a-b)*(c+d)} \tag{13}$$

6.2. Description of the GEP algorithm

The purpose of symbolic regression or function finding is to find an expression that can give a good explanation for the dependent variable. The first step is to choose the fitness function. Mathematically, the fitness f_i of an individual program i is expressed by the equation:

$$f_i = \sum_{j=1}^{n}\left(R - \left|\frac{P_{(ij)} - T_j}{T_j}\cdot 100\right|\right) \tag{14}$$

where R is the selection range, $P_{(ij)}$ is the value predicted by the individual program i for fitness case j (out of n fitness cases), and T_j is the target value for fitness case j. Note that the absolute value term corresponds to the relative error. This term is what is called the precision and if the error is smaller than or is equal to the precision then the error becomes zero. Thus, for a good match the absolute value term is zero and $f_i = f_{max} = nR$. For some function finding problems it is important to evolve a model that performs well for all fitness cases within a certain relative error (the precision) of the correct value:

$$f_{(i,j)} = \begin{cases} 1, E_{(i,j)} \le p \\ 0, \end{cases} \tag{15}$$

where p is the precision and $E_{(ij)}$ is the relative error of an individual program i for the fitting case j.

The $E_{(ij)}$ is given by:

$$E_{(ij)} = \left|\frac{P_{(ij)} - T_j}{T_j} \bullet 100\right| \tag{16}$$

The second step consists of choosing the set of terminals T and the set of functions F to create the chromosomes. In this problem, the terminal set consists obviously of the independent variable, i.e., $T = \{a\}$. The third step is to choose the chromosomal architecture, i.e., the length of the head and the number of genes. The fourth major step is to choose the linking function. The last major step is to choose the set of genetic operators that cause variation and their rates. These processes are repeated for a pre-specified number of generations until a solution is obtained. In the GEP, the individuals are often selected and copied into the next generation based on their fitness, as determined by roulette-wheel sampling with elitism, which guarantees the survival and cloning of the best individual to the next generation. The variation in the population is introduced by applying one or more genetic operators to select chromosomes, including crossover, mutation, and rotation. The process begins with the random generation of the chromosomes of the initial population. The chromosomes are expressed and the fitness of each individual is evaluated. The individuals are selected according to the fitness to reproduce with modification, leaving progeny with new traits. The individuals of this new generation are, in their turn, subjected to the same developmental process: expression of the genomes, confrontation of the selection environment, and reproduction with modification. To evaluate the ability of the GEP, the correlation coefficient (*R*) was introduced as:

$$C_i = \frac{\mathrm{Cov}(T,P)}{\sigma_t \sigma_p} \tag{17}$$

where Cov(*T*, *P*) is the covariance of the target and model outputs. σ_t and σ_p are the corresponding standard deviations. GEP is the newest chemometrics method, and Si *et al.* [23,25,19-121] have applied this method to QSAR studies for the first time. The results from their studies are satisfactory and show a promising use in the nonlinear structure-activity/property relationship correlation area, but GEP is congenitally defective as far as reproducibility of the predicted results is concerned, and always deduces too complex equations. This means a higher requirement for a user who is involved with GEP. GEP is now commercialized in the software Automatic Problem Solver 3.0 or GeneXproTools 4.0 [122].

7. Project Pursuit Regression (PPR)

PPR, which was developed by Friedman and Stuetzle [123], is a powerful tool for seeking the interesting projections from high-dimensional data into lower dimensional space by means of linear projections. Therefore, it can overcome the curse of dimensionality because it relies on estimation in at most trivariate settings. Friedman and Stuetzle's concept of PPR avoided many difficulties experienced with other existing nonparametric regression procedures. It does not split the predictor space into two regions, thereby allowing, when necessary, more complex models. In addition, interactions of predictor variables are directly considered because linear combinations of the predictors are modeled with general smooth functions. Another significant property of PPR is that the results of each interaction can be depicted graphically. The graphical output can be used to modify the major parameters of the procedure: the average smoother bandwidth and the terminal threshold. A brief description about PPR is presented here. Given the (k × n) data matrix X, where k is the number of observed variables and n is the number of units, and an m-dimensional orthonormal matrix A (m × k), the (m × n) matrix $Y = AX$ represents the coordinates of the projected data onto the m-dimensional ($m < k$) space spanned by the rows of A. Because such projections are infinite, it is important to have a technique to pursue a finite sequence of projections that can reveal the most interesting structures of the data. Projection pursuit (PP) is such a powerful tool that combines both ideas of projection and pursuit. In a typical regression problem, PPR aims to approximate the regression pursuit function $f(x)$ by a finite sum of ridge functions with suitable choices of α_i and g_i:

$$g^{(p)}(x) = \sum_{i=1}^{p} g_i(\alpha_i^T x) \tag{18}$$

where α_i values are m × n orthonormal matrices and p is the number of ridge functions. In recent QSAR/QSPR studies [21,37,124-129] PPR was employed as a regression method and always resulted in the best final models. This indicates that PPR is a promising regression method in QSAR/QSPR studies, especially when the correlation between descriptors and activities or properties is nonlinear.

8. Local Lazy Regression (LLR)

Most QSAR/QSPR models often capture the global structure-activity/property trends which are present in a whole dataset. In many cases, there may be groups of molecules which exhibit a specific set of features which relate to their activity or property. Such a major feature can be said to represent a local structure activity/property relationship. Traditional models may not recognize such local relationships. LLR is an excellent approach which extracts a prediction by locally interpolating the neighboring examples of the query which are considered relevant according to a distance measure, rather than considering the whole dataset. That will cause the basic core of this approach which is a simple assumption that similar compounds have similar activities or properties; that is, the activities or properties of molecules will change concurrently with the changes in the chemical structure. For one or more query points, "lazy" estimates the value of an unknown multivariate function on the basis of a set of possibly noisy samples of the function itself. Each sample is an input/output pair where the input is a vector and the output is a number. For each query point, the estimation of the input is obtained by combining different local models. Local models considered for combination by lazy are polynomials of zeroth, first, and second degree that fit a set of samples in the neighborhood of the query point. The neighbors are selected according to either the "Manhattan" or the "Euclidean" distance. It is possible to assign weights to the different directions of the input domain for modifying their importance in computation of the distance. The number of neighbors used for identifying local models is automatically adjusted on a query-by-query basis through a leave-one-out validation of models, each fitting a different number of neighbors. The local models are identified by using the recursive least-squares algorithm, and the leave-one-out cross-validation is obtained through the PRESS statistic. We assumed that there existed a small region around *x* which is linear between the dependent variable and the predictor variables. Then we determined the points around $X_{NN(x)}$ and build a regression model with only the points in $NN_{(x)}$ using the least-squares method and minimized the squared residuals for the region using this model. So we do not need to build multiple models for each point in the training set beforehand. In essence, when faced with a query point, the approach builds a representative predictive model. Hence, this approach is termed local LLR, which is described below:

$$\beta_x = (X^T_{NN(X)} X_{NN(x)})^{-1} X^T_{NN(X)} Y_{NN(x)} \tag{19}$$

where $X_{NN(x)}$ was a matrix of independent variables, $Y_{NN(x)}$ the column vectors representing the dependent variables, for the molecules in the neighborhood of the query point. β_x was the column vectors of regression coefficients. One of the main advantages of LLR is the fact that no *a priori* model needs to be built. This makes it suitable for large data sets, where using all of the observations can normally be time-consuming and even lead to over-fitting. At the same time, because it builds a regression model for each query point, one cannot extract meaningful structure-activity trends for the data set as a whole. That is, the focus of LLR is on predictive ability, rather than interpretability. Like every method, the lazy approach has a number of shortcomings. First, as all of the computations are done at query time, the determination of the local neighborhood must be efficient. Second, uncorrelated features might result in errors in the identification of near neighbors. Finally, it is nontrivial to integrate feature selection in this framework. LLR is generally used to develop linear models for data sets in which the global structure-activity/property relationship is nonlinear in nature.

However, as a new arising method in the field of QSAR/QSPR, LLR is not used extensively, with only a few relevant studies shown in references [125,130-132]. It is expected that more application studies involving LLR will appear in future QSAR/QSPR analyses.

9. Conclusions

In this paper, we focus on the current mathematical methods used as regression tools in recent QSAR/QSPR studies. Mathematical regression methods are so important for the QSAR/QSPR modeling that the choice of the regression method, most of the time, will determine if the resulted model will be successful or not. Fortunately, more and more new methods and algorithms have been applied to the studies of QSAR/QSPR, including linear and nonlinear, statistics and machine learning. At the same time, the existing methods have been improved. However, it is still a challenge for the researchers to choose suitable methods for modeling their systems. This paper may give some help on the knowledge of these methods, but more practical applications are needed so as to get a thorough understanding and then perform a better application.

References and Notes

1. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA) Ref. Man. Version 2.7.10, 2007.
2. Du, H.; Wang, J.; Hu, Z.; Yao, X. Quantitative Structure-Retention relationship study of the constituents of saffron aroma in SPME-GC-MS based on the projection pursuit regression method. *Talanta* **2008**, *77*, 360-365.
3. Du, H.; Watzl, J.; Wang, J.; Zhang, X.; Yao, X.; Hu, Z. Prediction of retention indices of drugs based on immobilized artificial membrane chromatography using Projection Pursuit Regression and Local Lazy Regression. *J. Sep. Sci.* **2008**, *31*, 2325-2333.
4. Du, H.; Zhang, X.; Wang, J.; Yao, X.; Hu, Z. Novel approaches to predict the retention of histidine-containing peptides in immobilized metal-affinity chromatography. *Proteomics* **2008**, *8*, 2185-2195.
5. Katritzky, A.R.; Pacureanu, L.; Dobchev, D.; Karelson, M. QSPR modeling of hyperpolarizabilities. *J. Mol. Model.* **2007**, *13*, 951-963.
6. Ren, Y.; Liu, H.; Yao, X.; Liu, M. An accurate QSRR model for the prediction of the GCxGC-TOFMS retention time of polychlorinated biphenyl (PCB) congeners. *Anal. Bioanal. Chem.* **2007**, *388*, 165-172.
7. Srivani, P.; Srinivas, E.; Raghu, R.; Sastry, G.N. Molecular modeling studies of pyridopurinone derivatives--potential phosphodiesterase 5 inhibitors. *J. Mol. Graph. Model.* **2007**, *26*, 378-390.
8. Kahn, I.; Sild, S.; Maran, U. Modeling the toxicity of chemicals to Tetrahymena pyriformis using heuristic multilinear regression and heuristic back-propagation neural networks. *J. Chem. Inf. Model.* **2007**, *47*, 2271-2279.
9. Semichem Home Page. Available online: http://www.semichem.com/codessa (accessed on 10 March 2009).

10. Codessa Pro Home Page. Available online: http://www.codessa-pro.com/ (accessed on 10 March 2009).
11. Xia, B.; Liu, K.; Gong, Z.; Zheng, B.; Zhang, X.; Fan, B. Rapid toxicity prediction of organic chemicals to Chlorella vulgaris using quantitative structure-activity relationships methods. *Ecotoxicol. Environ. Saf.* **2009**, *72*, 787-794.
12. Yuan, Y.; Zhang, R.; Hu, R.; Ruan, X. Prediction of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas based on the heuristic method, support vector machine and projection pursuit regression. *Eur. J. Med. Chem.* **2009**, *44*, 25-34.
13. Lu, W.J.; Chen, Y.L.; Ma, W.P.; Zhang, X.Y.; Luan, F.; Liu, M.C.; Chen, X.G.; Hu, Z.D. QSAR study of neuraminidase inhibitors based on heuristic method and radial basis function network. *Eur. J. Med. Chem.* **2008**, *43*, 569-576.
14. Xia, B.; Ma, W.; Zheng, B.; Zhang, X.; Fan, B. Quantitative structure-activity relationship studies of a series of non-benzodiazepine structural ligands binding to benzodiazepine receptor. *Eur. J. Med. Chem.* **2008**, *43*, 1489-1498.
15. Zhao, C.; Zhang, H.; Luan, F.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. QSAR method for prediction of protein-peptide binding affinity: application to MHC class I molecule HLA-A*0201. *J. Mol. Graph. Model.* **2007**, *26*, 246-254.
16. Rebehmed, J.; Barbault, F.; Teixeira, C.; Maurel, F. 2D and 3D QSAR studies of diarylpyrimidine HIV-1 reverse transcriptase inhibitors. *J. Comput. Aided Mol. Des.* **2008**, *22*, 831-841.
17. Agrafiotis, D.K.; Gibbs, A.C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.* **2007**, *47*, 1067-1086.
18. Si, H.Z.; Wang, T.; Zhang, K.J.; Hu, Z.D.; Fan, B.T. QSAR study of 1,4-dihydropyridine calcium channel antagonists based on gene expression programming. *Bioorg. Med. Chem.* **2006**, *14*, 4834-4841.
19. Li, X.; Luan, F.; Si, H.; Hu, Z.; Liu, M. Prediction of retention times for a large set of pesticides or toxicants based on support vector machine and the heuristic method. *Toxicol. Lett.* **2007**, *175*, 136-144.
20. Gong, Z.G.; Zhang, R.S.; Xia, B.B.; Hu, R.J.; Fan, B.T. Study of nematic transition temperatures in themotropic liquid crystal using heuristic method and radial basis function neural networks and support vector machine. *QSAR Comb.Sci.* **2008**, *27*, 1282-1290.
21. Yuan, Y.N.; Zhang, R.S.; Hu, R.J.; Ruan, X.F. Prediction of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas based on the heuristic method, support vector machine and projection pursuit regression. *Eur. J. Med. Chem.* **2009**, *44*, 25-34.
22. Xia, B.B.; Liu, K.P.; Gong, Z.G.; Zheng, B.; Zhang, X.Y.; Fan, B.T. Rapid toxicity prediction of organic chemicals to Chlorella vulgaris using quantitative structure-activity relationships methods. *Ecotoxicol. Environ. Saf.* **2009**, *72*, 787-794.
23. Luan, F.; Si, H.Z.; Liu, H.T.; Wen, Y.Y.; Zhang, X.Y. Prediction of atmospheric degradation data for POPs by gene expression programming. *SAR QSAR Environ. Res.* **2008**, *19*, 465-479.
24. Xia, B.B.; Ma, W.P.; Zheng, B.; Zhang, X.Y.; Fan, B.T. Quantitative structure-activity relationship studies of a series of non-benzodiazepine structural ligands binding to benzodiazepine receptor. *Eur. J. Med. Chem.* **2008**, *43*, 1489-1498.

25. Wang, T.; Si, H.Z.; Chen, P.P.; Zhang, K.J.; Yao, X.J. QSAR models for the dermal penetration of polycyclic aromatic hydrocarbons based on Gene Expression Programming. *QSAR Comb. Sci.* **2008**, *27*, 913-921.
26. Liu, K.P.; Xia, B.B.; Ma, W.P.; Zheng, B.; Zhang, X.Y.; Fan, B.T. Quantitative structure-activity relationship modeling of triaminotriazine drugs based on Heuristic Method. *QSAR Comb. Sci.* **2008**, *27*, 425-431.
27. Lu, W.J.; Chen, Y.L.; Ma, W.P.; Zhang, X.Y.; Luan, F.; Liu, M.C.; Chen, X.G.; Hu, Z.D. QSAR study of neuraminidase inhibitors based on heuristic method and radial basis function network. *Eur. J. Med. Chem.* **2008**, *43*, 569-576.
28. Zhao, C.Y.; Zhang, H.X.; Luan, F.; Zhang, R.S.; Liu, M.C.; Hu, Z.D.; Fan, B.T. QSAR method for prediction of protein-peptide binding affinity: Application to MHC class I molecule HLA-A*0201. *J. Mol. Graph. Model.* **2007**, *26*, 246-254.
29. Li, H.Z.; Liu, H.X.; Yao, X.J.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Quantitative structure-activity relationship study of acyl ureas as inhibitors of human liver glycogen phosphorylase using least squares support vector machines. *Chemometr. Intel. Lab. Syst.* **2007**, *87*, 139-146.
30. Qin, S.; Liu, H.X.; Wang, J.; Yao, X.J.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Quantitative Structure-Activity Relationship study on a series of novel ligands binding to central benzodiazepine receptor by using the combination of Heuristic Method and Support Vector Machines. *QSAR Comb. Sci.* **2007**, *26*, 443-451.
31. Ma, W.P.; Luan, F.; Zhao, C.Y.; Zhang, X.Y.; Liu, M.C.; Hu, Z.D.; Fan, B.T. QSAR prediction of the penetration of drugs across a polydimethylsiloxane membrane. *QSAR Comb. Sci.* **2006**, *25*, 895-904.
32. Luan, F.; Ma, W.P.; Zhang, X.Y.; Zhang, H.X.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Quantitative structure-activity relationship models for prediction of sensory irritants (logRD(50)) of volatile organic chemicals. *Chemosphere* **2006**, *63*, 1142-1153.
33. Si, H.Z.; Yao, X.J.; Liu, H.X.; Wang, J.; Li, J.Z.; Hu, Z.D.; Liu, M.C. Prediction of binding rate of drug to human plasma protein based on heuristic method and support vector machine. *Acta Chim. Sinica* **2006**, *64*, 415-422.
34. Luan, F.; Ma, W.P.; Zhang, X.Y.; Zhang, H.X.; Liu, M.C.; Hu, Z.D.; Fan, B.T. QSAR study of polychlorinated dibenzodioxins, dibenzofurans, and Biphenyls using the heuristic method and support vector machine. *QSAR Comb. Sci.* **2006**, *25*, 46-55.
35. Gharagheizi, F.; Tirandazi, B.; Barzin, R. Estimation of aniline point temperature of pure hydrocarbons: A quantitative structure-property relationship approach. *Ind. Eng. Chem. Res.* **2009**, *48*, 1678-1682.
36. Riahi, S.; Mousavi, M.F.; Ganjali, M.R.; Norouzi, P. Application of correlation ranking procedure and artificial neural networks in the modeling of liquid chromatographic retention times (tR) of various pesticides. *Anal. Lett.* **2008**, *41*, 3364-3385.
37. Du, H.Y.; Wang, J.; Hu, Z.D.; Yao, X.J.; Zhang, X.Y. Prediction of fungicidal activities of rice blast disease based on least-squares support vector machines and project pursuit regression. *J. Agric. Food Chem.* **2008**, *56*, 10785-10792.
38. Gharagheizi, F.; Mehrpooya, M. Prediction of some important physical properties of sulfur compounds using quantitative structure-properties relationships. *Mol. Div.* **2008**, *12*, 143-155.

39. Sattari, M.; Gharagheizi, F. Prediction of molecular diffusivity of pure components into air: A QSPR approach. *Chemosphere* **2008**, *72*, 1298-1302.
40. Gharagheizi, F.; Alamdari, R.F. Prediction of flash point temperature of pure components using a Quantitative Structure-Property Relationship model. *QSAR Comb. Sci.* **2008**, *27*, 679-683.
41. Gharagheizi, F.; Fazeli, A. Prediction of the Watson characterization factor of hydrocarbon components from molecular properties. *QSAR Comb. Sci.* **2008**, *27*, 758-767.
42. Om, A.S.; Ryu, J.C.; Kim, J.H. Quantitative structure-activity relationships for radical scavenging activities of flavonoid compounds by GA-MLR technique. *Mol. Cell. Toxicol.* **2008**, *4*, 170-176.
43. Riahi, S.; Ganjali, M.R.; Pourbasheer, E.; Norouzi, P. QSRR study of GC retention indices of essential-oil compounds by multiple linear regression with a genetic algorithm. *Chromatographia* **2008**, *67*, 917-922.
44. Hashemianzadeh, M.; Safarpour, M.A.; Gholamjani-Moghaddam, K.; Mehdipour, A.R. DFT-based QSAR study of valproic acid and its derivatives. *QSAR Comb. Sci.* **2008**, *27*, 469-474.
45. Gharagheizi, F. A new molecular-based model for prediction of enthalpy of sublimation of pure components. *Thermochim. Acta* **2008**, *469*, 8-11.
46. Gharagheizi, F. QSPR studies for solubility parameter by means of Genetic Algorithm-Based Multivariate Linear Regression and generalized regression neural network. *QSAR Comb. Sci.* **2008**, *27*, 165-170.
47. Gharagheizi, F.; Alamdari, R.F. A molecular-based model for prediction of solubility of C-60 fullerene in various solvents. *Fuller. Nanotub. Carbon Nanostr.* **2008**, *16*, 40-57.
48. Carlucci, G.; D'Archivio, A.A.; Maggi, M.A.; Mazzeo, P.; Ruggieri, F. Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-performance liquid chromatography by using quantitative structure-retention relationships. *Anal. Chim. Acta* **2007**, *601*, 68-76.
49. Gharagheizi, F. A new accurate neural network quantitative structure-property relationship for prediction of theta (lower critical solution temperature) of polymer solutions. *E-Polymers* **2007**, No. 114.
50. Elliott, G.N.; Worgan, H.; Broadhurst, D.; Draper, J.; Scullion, J. Soil differentiation using fingerprint Fourier transform infrared spectroscopy, chemometrics and genetic algorithm-based feature selection. *Soil Biol. Biochem.* **2007**, *39*, 2888-2896.
51. Gharagheizi, F. QSPR analysis for intrinsic viscosity of polymer solutions by means of GA-MLR and RBFNN. *Comput. Mater. Sci.* **2007**, *40*, 159-167.
52. Deeb, O.; Hemmateenejad, B.; Jaber, A.; Garduno-Juarez, R.; Miri, R. Effect of the electronic and physicochemical parameters on the carcinogenesis activity of some sulfa drugs using QSAR analysis based on genetic-MLR and genetic-PLS. *Chemosphere* **2007**, *67*, 2122-2130.
53. Vatani, A.; Mehrpooya, M.; Gharagheizi, F. Prediction of standard enthalpy of formation by a QSPR model. *Int. J. Mol. Sci.* **2007**, *8*, 407-432.
54. Jung, M.; Tak, J.; Lee, Y.; Jung, Y. Quantitative structure-activity relationship (QSAR) of tacrine derivatives against acetylcholinesterase (ACNE) activity using variable selections. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1082-1090.

55. Fisz, J.J. Combined genetic algorithm and multiple linear regression (GA-MLR) optimizer: Application to multi-exponential fluorescence decay surface. *J. Phys. Chem. A* **2006**, *110*, 12977-12985.
56. Word, H. *Research Papers in Statistics*; Wiley: New York, NY, USA, 1966.
57. Jores-Kong, H.; Word, H. *Systems under Indirect Observation: Causality, structure, prediction*; North-Holland: Amsterdam, The Netherlands, 1982.
58. Rogers, D.; Hopfinger, A.J. Application of genetic function approximation to quantitative structure-activity-relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854-866.
59. Fan, Y.; Shi, L.M.; Kohn, K.W.; Pommier, Y.; Weinstein, J.N. Quantitative structure-antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies. *J. Med. Chem.* **2001**, *44*, 3254-3263.
60. Sammi, T.; Silakari, O.; Ravikumar, M. Three-dimensional quantitative structure-activity relationship (3D-QSAR) studies of various benzodiazepine analogues of gamma-secretase inhibitors. *J. Mol. Model.* **2009**, *15*, 343-348.
61. Li, Z.G.; Chen, K.X.; Xie, H.Y.; Gao, J.R. Quantitative structure - activity relationship analysis of some thiourea derivatives with activities against HIV-1 (IIIB). *QSAR Comb. Sci.* **2009**, *28*, 89-97.
62. Samee, W.; Nunthanavanit, P.; Ungwitayatorn, J. 3D-QSAR investigation of synthetic antioxidant chromone derivatives by molecular field analysis. *Int. J. Mol. Sci.* **2008**, *9*, 235-246.
63. Nunthanavanit, P.; Anthony, N.G.; Johnston, B.F.; Mackay, S.P.; Ungwitayatorn, J. 3D-QSAR studies on chromone derivatives as HIV-1 protease inhibitors: Application of molecular field analysis. *Arch. Der. Pharm.* **2008**, *341*, 357-364.
64. Kansal, N.; Silakari, O.; Ravikumar, M. 3D-QSAR studies of various diaryl urea derivatives of multi-targeted receptor tyrosine kinase inhibitors: Molecular field analysis approach. *Lett. Drug Des. Dis.* **2008**, *5*, 437-448.
65. Joseph, T.B.; Kumar, B.; Santhosh, B.; Kriti, S.; Pramod, A.B.; Ravikumar, M.; Kishore, M. Quantitative structure activity relationship and pharmacophore studies of adenosine receptor A(2B) inhibitors. *Chem. Biol. Drug Des.* **2008**, *72*, 395-408.
66. Equbal, T.; Silakari, O.; Ravikumar, M. Exploring three-dimensional quantitative structural activity relationship (3D-QSAR) analysis of SCH 66336 (Sarasar) analogues of farnesyltransferase inhibitors. *Eur. J. Med. Chem.* **2008**, *43*, 204-209.
67. Bhonsle, J.B.; Bhattacharjee, A.K.; Gupta, R.K. Novel semi-automated methodology for developing highly predictive QSAR models: application for development of QSAR models for insect repellent amides. *J. Mol. Model.* **2007**, *13*, 179-208.
68. Thomas Leonard, J.; Roy, K. Comparative QSAR modeling of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4467-4474.
69. Roy, K.; Leonard, J.T. Topological QSAR modeling of cytotoxicity data of anti-HIV 5-phenyl-1-phenylamino-imidazole derivatives using GFA, G/PLS, FA and PCRA techniques. *Indian J. Chem. Sect. A-Inorg. Bio-Inorg. Phys. Theor. Anal. Chem.* **2006**, *45*, 126-137.

70. Davies, M.N.; Hattotuwagama, C.K.; Moss, D.S.; Drew, M.G.B.; Flower, D.R. Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity. *BMC Struct. Biol.* **2006**, *6*, 5:1-5:13.
71. Mandal, A.S.; Roy, K. Predictive QSAR modeling of HIV reverse transcriptase inhibitor TIBO derivatives. *Eur. J. Med. Chem.* **2009**, *44*, 1509-1524.
72. Leonard, J.T.; Roy, K. Exploring molecular shape analysis of styrylquinoline derivatives as HIV-1 integrase inhibitors. *Eur. J. Med. Chem.* **2008**, *43*, 81-92.
73. Roy, K.; Ghosh, G. QSTR with extended topochemical atom (ETA) indices 8.(a) QSAR for the inhibition of substituted phenols on germination rate of Cucumis sativus using chemometric tools. *QSAR Comb. Sci.* **2006**, *25*, 846-859.
74. Leonard, J.T.; Roy, K. Comparative QSAR modeling of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4467-4474.
75. Wold, S.; Antti, H.; Lindgren, F.; Ohman, J. Orthogonal signal correction of near-infrared spectra. *Chemometr. Intel. Lab. Syst.* **1998**, *44*, 175-185.
76. Yin, P.Y.; Mohemaiti, P.; Chen, J.; Zhao, X.J.; Lu, X.; Yimiti, A.; Upur, H.; Xu, G.W. Serum metabolic profiling of abnormal savda by liquid chromatography/mass spectrometry. *J. Chromatogr. B-Anal. Technol. Biomed. Life Sci.* **2008**, *871*, 322-327.
77. Samadi-Maybodi, A.; Darzi, S. Simultaneous determination of vitamin B12 and its derivatives using some of multivariate calibration 1 (MVC1) techniques. *Spectrochim. Acta A-Mol. Biomol. Spectrosc.* **2008**, *70*, 1167-1172.
78. Niazi, A.; Jafarian, B.; Ghasemi, J. Kinetic spectrophotometric determination of trace amounts of palladium by whole kinetic curve and a fixed time method using resazurine sulfide reaction. *Spectrochim. Acta A-Mol. Biomol. Spectrosc.* **2008**, *71*, 841-846.
79. Niazi, A.; Goodarzi, M. Orthogonal signal correction-partial least squares method for simultaneous spectrophotometric determination of cypermethrin and tetramethrin. *Spectrochim. Acta A-Mol. Biomol. Spectrosc.* **2008**, *69*, 1165-1169.
80. Niazi, A.; Amjadi, E.; Nori-Shargh, D.; Bozorghi, S.J. Simultaneous voltammetric determination of lead and tin by adsorptive differential pulse stripping method and orthogonal signal correction-partial least squares in water samples. *J. Chinese Chem. Soc.* **2008**, *55*, 276-285.
81. Karimi, M.A.; Ardakani, M.M.; Behjatmanesh-Ardakani, R.; Nezhad, M.R.H.; Amiryan, H. Individual and simultaneous determinations of phenothiazine drugs using PCR, PLS and (OSC)-PLS multivariate calibration methods. *J. Serb. Chem. Soc.* **2008**, *73*, 233-247.
82. Cho, H.W.; Kim, S.B.; Jeong, M.K.; Park, Y.; Miller, N.G.; Ziegler, T.R.; Jones, D.P. Discovery of metabolite features for the modelling and analysis of high-resolution NMR spectra. *Int. J. Data Min. Bioinf.* **2008**, *2*, 176-192.
83. Cheng, Z.; Zhu, A.S.; Zhang, L.Q. Quantitative analysis of electronic absorption spectroscopy by piecewise orthogonal signal correction and partial least square. *Guang Pu Xue Yu Guang Pu Fen Xi* **2008**, *28*, 860-864.
84. Cheng, Z.; Zhu, A.S.; Zhang, L.Q. Quantitative analysis of electronic absorption spectroscopy by piecewise orthogonal signal correction and partial least square. *Spectrosc. Spectr. Anal.* **2008**, *28*, 860-864.

85. Cheng, Z.; Zhu, A.S. Piecewise orthogonal signal correction approach and its application in the analysis of wheat near-infrared spectroscopic data. *Chinese J. Anal. Chem.* **2008**, *36*, 788-792.
86. Rouhollahi, A.; Rajabzadeh, R.; Ghasemi, J. Simultaneous determination of dopamine and ascorbic acid by linear sweep voltammetry along with chemometrics using a glassy carbon electrode. *Microchim. Acta* **2007**, *157*, 139-147.
87. Psihogios, N.G.; Kalaitzidis, R.G.; Dimou, S.; Seferiadis, K.I.; Siamopoulos, K.C.; Bairaktari, E.T. Evaluation of tubulointerstitial lesions' severity in patients with glomerulonephritides: An NMR-Based metabonomic study. *J. Proteome Res.* **2007**, *6*, 3760-3770.
88. Niazi, A.; Azizi, A.; Leardi, R. A comparative study between PLS and OSC-PLS in the simultaneous determination of lead and mercury in water samples: effect of wavelength selection. *Can. J. Anal. Sci. Spectrosc.* **2007**, *52*, 365-374.
89. Priolo, N.; Arribere, C.M.; Caffini, N.; Barberis, S.; Vazquez, R.N.; Luco, J.M. Isolation and purification of cysteine peptidases from the latex of Araujia hortorum fruits - Study of their esterase activities using partial least-squares (PLS) modeling. *J. Mol. Catal. B-Enzym.* **2001**, *15*, 177-189.
90. Yang, S.B.; Xia, Z.N.; Shu, M.; Mei, H.; Lue, F.L.; Zhang, M.; Wu, Y.Q.; Li, Z.L. VHSEH Descriptors for the Development of QSAMs of Peptides. *Chem. J. Chinese Univ.* **2008**, *29*, 2213-2217.
91. Liang, G.Z.; Mei, H.; Zhou, Y.; Yang, S.B.; Wu, S.R.; Li, Z.L. Using SZOTT descriptors for the development of QSAMs of peptides. *Chem. J. Chinese Univ.* **2006**, *27*, 1900-1902.
92. Zhao, C.Y.; Boriani, E.; Chana, A.; Roncaglioni, A.; Benfenati, E. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* **2008**, *73*, 1701-1707.
93. Qi, J.; Niu, J.F.; Wang, L.L. Research on QSPR for n-octanol-water partition coefficients of organic compounds based on genetic algorithms-support vector machine and genetic algorithms-radial basis function neural networks. *Huanjing Kexue* **2008**, *29*, 212-218.
94. Luan, F.; Liu, H.T.; Wen, Y.Y.; Zhang, X.Y. Prediction of quantitative calibration factors of some organic compounds in gas chromatography. *Analyst* **2008**, *133*, 881-887.
95. Luan, F.; Liu, H.T.; Wen, Y.Y.; Zhang, X.Y. Quantitative structure-property relationship study for estimation of quantitative calibration factors of some organic compounds in gas chromatography. *Anal. Chim. Acta* **2008**, *612*, 126-135.
96. Luan, F.; Liu, H.T.; Ma, W.P.; Fan, B.T. QSPR analysis of air-to-blood distribution of volatile organic compounds. *Ecotoxicol. Environ. Saf.* **2008**, *71*, 731-739.
97. Chen, H.F. Quantitative predictions of gas chromatography retention indexes with support vector machines, radial basis neural networks and multiple linear regression. *Anal. Chim. Acta* **2008**, *609*, 24-36.
98. Zhao, C.Y.; Zhang, H.X.; Zhang, X.Y.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology* **2006**, *217*, 105-119.
99. Tetko, I.V.; Solov'ev, V.P.; Antonov, A.V.; Yao, X.J.; Doucet, J.P.; Fan, B.T.; Hoonakker, F.; Fourches, D.; Jost, P.; Lachiche, N.; Varnek, A. Benchmarking of linear and nonlinear approaches for quantitative structure-property relationship studies of metal complexation with ionophores. *J. Chem. Inf. Model.* **2006**, *46*, 808-819.

100. Shi, J.; Luan, F.; Zhang, H.X.; Liu, M.C.; Guo, Q.X.; Hu, Z.D.; Fan, B.T. QSPR study of fluorescence wavelengths (lambda(ex)/lambda(em)) based on the heuristic method and radial basis function neural networks. *QSAR Comb. Sci.* **2006**, *25*, 147-155.
101. Ma, W.P.; Luan, F.; Zhang, H.X.; Zhang, X.Y.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Accurate quantitative structure-property relationship model of mobilities of peptides in capillary zone electrophoresis. *Analyst* **2006**, *131*, 1254-1260.
102. Luan, F.; Zhang, X.; Zhang, H.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. Prediction of standard Gibbs energies of the transfer of peptide anions from aqueous solution to nitrobenzene based on support vector machine and the heuristic method. *J. Comput. Aided Mol. Des.* **2006**, *20*, 1-11.
103. Specht, D.F. A general regression neural network. *IEEE Trans. Neur. Netw.* **1991**, *2*, 568-576.
104. Szaleniec, M.; Tadeusiewicz, R.; Witko, M. How to select an optimal neural model of chemical reactivity? *Neurocomputing* **2008**, *72*, 241-256.
105. Ji, L.; Wang, X.D.; Luo, S.; Qin, L.A.; Yang, X.S.; Liu, S.S.; Wang, L.S. QSAR study on estrogenic activity of structurally diverse compounds using generalized regression neural network. *Sci. China Ser. B-Chem.* **2008**, *51*, 677-683.
106. Mager, P.P. Subset selection and docking of human P2X7 inhibitors. *Curr. Comput. Aided Drug Des.* **2007**, *3*, 248-253.
107. Ibric, S.; Jovanovic, M.; Djuric, Z.; Parojcic, J.; Solomun, L.; Lucic, B. Generalized regression neural networks in prediction of drug stability. *J. Pharm. Pharmacol.* **2007**, *59*, 745-750.
108. Yap, C.W.; Li, Z.R.; Chen, Y.Z. Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. *J. Mol. Graph. Model.* **2006**, *24*, 383-395.
109. Agatonovic-Kustrin, S.; Turner, J.V. Artificial neural network modeling of phytoestrogen binding to estrogen receptors. *Lett. Drug Des. Disc.* **2006**, *3*, 436-442.
110. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273-297.
111. Vapnik, V. The Support Vector method of function estimation. *U.S. Patent 5,950,146*, 1999.
112. Wang, W.J.; Xu, Z.B.; Lu, W.Z.; Zhang, X.Y. Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing* **2003**, *55*, PII S0925-2312(02)00632-X.
113. Suykens, J.A.K.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293-300.
114. Yuan, Y.N.; Zhang, R.S.; Hu, R.J.; Ruan, X.F. Prediction of volatile components retention time in blackstrap molasses by least-squares support vector machine. *QSAR Comb. Sci.* **2008**, *27*, 535-542.
115. Niazi, A.; Jameh-Bozorghi, S.; Nori-Shargh, D. Prediction of toxicity of nitrobenzenes using ab initio and least squares support vector machines. *J. Hazard. Mater.* **2008**, *151*, 603-609.
116. Goudarzi, N.; Goodarzi, M. Prediction of the logarithmic of partition coefficients (log P) of some organic compounds by least square-support vector machine (LS-SVM). *Mol. Phys.* **2008**, *106*, DOI 10.1080/00268970802577834|PII 905993443.
117. Liu, H.; Papa, E.; Walker, J.D.; Gramatica, P. In silico screening of estrogen-like chemicals based on different nonlinear classification models. *J. Mol. Graph. Model.* **2007**, *26*, 135-144.

118. Liu, H.X.; Yao, X.J.; Zhang, R.S.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Prediction of the tissue/blood partition coefficients of organic compounds based on the molecular structure using least-squares support vector machines. *J. Comput. Aided Mol. Des.* **2005**, *19*, 499-508.
119. Si, H.Z.; Wang, T.; Zhang, K.J.; Duan, Y.B.; Yuan, S.P.; Fu, A.P.; Hu, Z.D. Quantitative structure activity relationship model for predicting the depletion percentage of skin allergic chemical substances of glutathione. *Anal. Chim. Acta* **2007**, *591*, 255-264.
120. Si, H.Z.; Zhang, K.J.; Hu, Z.D.; Fan, B.T. QSAR model for prediction capacity factor of molecular imprinting polymer based on gene expression programming. *QSAR Comb. Sci.* **2007**, *26*, 41-50.
121. Si, H.Z.; Yuan, S.P.; Zhang, K.J.; Fu, A.P.; Duan, Y.B.; Hue, Z.D. Quantitative structure activity relationship study on EC5.0 of anti-HIV drugs. *Chemometr. Intel. Lab. Syst.* **2008**, *90*, 15-24.
122. Gepsoft Home Page. Available online: http://www.gepsoft.com/ (accessed on 10 March 2009).
123. Friedman, J.H.; Stuetzle, W. Projection Pursuit Regression. *J. Am. Stat. Assoc.* **1981**, *76*, 817-823.
124. Yuan, Y.N.; Zhang, R.S.; Hu, R.J. Prediction of Photolysis of PCDD/Fs Adsorbed to Spruce [Picea abies (L.) Karst.] Needle Surfaces Under Sunlight Irradiation Based on Projection Pursuit Regression. *QSAR Comb. Sci.* **2009**, *28*, 155-162.
125. Du, H.Y.; Zhang, X.Y.; Wang, X.; Yao, X.J.; Hu, Z.D. Novel approaches to predict the retention of histidine-containing peptides in immobilized metal-affinity chromatography. *Proteomics* **2008**, *8*, 2185-2195.
126. Du, H.Y.; Wang, J.; Zhang, X.Y.; Hu, Z.D. A novel quantitative structure-activity relationship method to predict the affinities of MT3 melatonin binding site. *Eur. J. Med. Chem.* **2008**, *43*, 2861-2869.
127. Du, H.Y.; Wang, J.; Watzl, J.; Zhang, X.Y.; Hu, Z.D. Prediction of inhibition of matrix metalloproteinase inhibitors based on the combination of Projection Pursuit Regression and Grid Search method. *Chemometr. Intel. Lab. Syst.* **2008**, *93*, 160-166.
128. Ren, Y.Y.; Liu, H.X.; Yao, X.J.; Liu, M.C. Prediction of ozone tropospheric degradation rate constants by projection pursuit regression. *Anal. Chim. Acta* **2007**, *589*, 150-158.
129. Ren, Y.Y.; Liu, H.X.; Li, S.Y.; Yao, X.J.; Liu, M.C. Prediction of binding affinities to beta(1) isoform of human thyroid hormone receptor by genetic algorithm and projection pursuit regression. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2474-2482.
130. Gunturi, S.B.; Archana, K.; Khandelwal, A.; Narayanan, R. Prediction of hERG Potassium Channel Blockade Using kNN-QSAR and Local Lazy Regression Methods. *QSAR Comb. Sci.* **2008**, *27*, 1305-1317.
131. Du, H.Y.; Watzl, J.; Wang, J.; Zhang, X.Y.; Yaol, X.J.; Hu, Z.D. Prediction of retention indices of drugs based on immobilized artificial membrane chromatography using Projection Pursuit Regression and Local Lazy Regression. *J. Sep. Sci.* **2008**, *31*, 2325-2333.

132. Guha, R.; Dutta, D.; Jurs, P.C.; Chen, T. Local lazy regression: Making use of the neighborhood to improve QSAR predictions. *J. Chem. Inf. Model.* **2006**, *46*, 1836-1847.