VSMP: A Novel Variable Selection and Modeling Method Based on the Prediction

Shu-Shen Liu,*,†,‡ Hai-Ling Liu,‡ Chun-Sheng Yin,† and Lian-Sheng Wang†

State Key Laboratory of Pollution Control and Resources Reuse, Department of Environmental Science & Engineering, Nanjing University, Nanjing 210093, P. R. China, and Department of Applied Chemistry, Guilin Institute of Technology, Guilin 541004, P. R. China

Received July 8, 2002

The use of numerous descriptors that are indicative of molecular structure and topology is becoming more common in quantitative structure−activity relationship (QSAR). How to choose the adequate descriptors for QSAR studies is important but difficult because there are no absolute rules to govern this choice. A variety of variable selection techniques including stepwise, partial least squares/principal component analysis (PLS/PCA), neural network, and evolutionary algorithm such as genetic algorithm have been applied to this common problem. All-subsets regression (ASR) is capable of finding out the best variable subset from among a large pool. In this paper, a novel variable selection and modeling method based on the prediction, for short VSMP, has been developed. Here two controllable parameters, the interrelation coefficient between the pairs of the independent variables (r_{int}) and the correlation coefficient (q^2) obtained using the leave-one-out (LOO) cross-validation technique, are introduced into the ASR to improve its performances. This technique differs from the other variable selection procedures related to the ASR by two main features: (1) The search of various optimal subset search is controlled by the statistic q^2 or root-mean-square error (RMSEP) in the LOO cross-validation step rather than the correlation coefficient obtained in the modeling step (r^2). (2) The searching speed of all optimal subsets is expedited by the statistic r_{int} together with q^2. A comparison of the results of the VSMP applied to the Selwood data set ($n = 31$ compounds, $m = 53$ descriptors) with those obtained from alternative algorithms shows the good performance of the technique.

INTRODUCTION

Building a useful quantitative structure−activity relationship (QSAR) model is an indispensable and important work in modern drug design. The QSAR approach usually includes the following successive steps: data preparation, data reduction, data modeling, and prediction. The success of a QSAR study relies heavily on how each of these steps is conducted and how the analysis is performed.[1] At present, the use of numerous descriptors that are indicative of molecular structure and topology is becoming more important for use in QSAR.[2] These types of descriptors are easily calculated from molecular structures and potentially number in the thousands. So, how to choose the adequate descriptors for QSAR studies has become very important. However, it is a difficult work because there are no absolute rules to govern this choice. To deal with this issue, a variety of variable selection techniques have been introduced.

The variable selection techniques based on stepwise regression[3−5] and partial least squares/principal component analysis (PLS/PCA)[6−8] are ones of most widespread used. The stepwise regression variable selection seems to be suitable for a little descriptor pool. PLS/PCA is regularly used to reduce a large descriptor pool to a manageable handful of latent variables related to the actual descriptors by a loafing matrix and applied in 3D-QSAR approaches such as notable CoMFA[9] and SOMFA[10] and 2D-QSAR ones such as HQSAR.[11] However, the actual physical meaning of those latent variables (i.e. principal components) is difficult to represent. On the other hand, the difference between the correlation coefficient in modeling step (r^2) and the cross-validation correlation coefficient (q^2) varies widely with increasing variables. More recently, some evolutionary algorithms[12] and neural network algorithms[13−15] were used for the variable selection problem. Genetic algorithms (GAs)[16−19] are very likely to be the most widely known type of evolutionary algorithms. Because of their simplicity, flexibility, easy operation, minimal requirements, and global perspective, GAs have been successfully used in many scientific fields. It is well-known that classical GA procedure is sometimes located into a localized optimum area to miss the best value. To settle the issue, GA program is run in various different initial populations but thus prolongs the computing time. All-subsets regression (ASR) [20−22] approach is a systematic searching method that can find out the best subset from a large descriptor pool but takes a lot of the computing time. Besides, the ASR procedure is in general based on the statistic r^2 rather than q^2. If the cross-validation technique is introduced into the ASR, the computing time is rapidly increased.

Golbraikh and Tropshathe[23] indicated that the high value of LOO q^2 appears to be the necessary but not the sufficient condition for the model having a high predictive power. Our previous results[24] also showed that a model with a high predictive power has to possess both the high values of LOO q^2 and r^2. So, in this paper, two important statistic parameters, the interrelated coefficient (r_{int}) between various pairs of the

* Corresponding author phone: (86)-025−3596509; e-mail: ssliuhl@263.net or ssliu@nju.edu.cn.
† Nanjing University.
‡ Guilin Institute of Technology.

10.1021/ci020377j CCC: $25.00
Published on Web 04/04/2003

independent variables and q^2, were introduced to accelerate the ASR course and obtain the optimal subset which has highest q^2 or lowest root-mean-square error (RMSEP) in the LOO cross-validation step and enough good r^2 or low root-mean-square error (RMSEE) in modeling. Unlike most current variable selection methods, in this paper, the determination of the best subset is based on q^2 predicted in the cross-validation prediction process rather than r^2 in modeling estimation process. So, the method is called the variable selection and modeling based on the prediction (for short VSMP).

METHODS

To accelerate the running speed of the classical all-subsets regression (ASR) and obtain the best variable subset evaluated by the predictive quality, two statistic parameters, one for the interrelated coefficient (r_{int}) between the variables and another for the LOO cross-validation correlation coefficient (q^2), are introduced into the ASR procedure to construct a novel computer program for the variable selection and modeling. It is assumed that the original data set consists of an independent variable matrix, $\boldsymbol{x}(n,m)$, which includes $n \times m$ descriptors of n compounds, and a dependent variable matrix, $\boldsymbol{y}(n)$, including n properties (biological activities or physiochemical properties) of the compounds. Then, how to select and analyze the best variable subset from among the original data set? Two main steps are designed to search the best subset in our present paper.

Selection of Various Optimal Subsets. First, an optimal subset is selected for a given number of variables (υn). This optimal subset is the best one for the given υn but not always the best for the whole subset space from m descriptors. The selective procedure of the optimal subset for the given υn is shown in Figure 1. The selective steps are as follows.

(1) The values of υn and r_{int} are specified where r_{int} is the allowable maximal threshold of the interrelated coefficient between various pairs of the independent variables. In this paper, let υn equal from 2, 3, 4, ..., until an integer of $n/5$ and $r_{int} = 0.76$. Here the r_{int} value of 0.76 replacing 0.75 is compared with the results in the literature[2] where the interrelated coefficient between a pair of the variables is 0.7539.

(2) The initial values of two important iterative statistics, r_{cri} and f_{max}, are specified. Here, the former, r_{cri}, is a control parameter used to determine whether the next LOO cross-validation step is run or not. The latter, f_{max}, is the maximal q^2 obtained in the previous loop. It determines the starting point in the next loop. The initial values of the r_{cri} and f_{max} cannot be larger than the final optimal value of q^2. For example, if the optimal value of q^2 is 0.70, the r_{cri} and f_{max} values of <0.70 are appropriate.

(3) A subset, $\boldsymbol{x}(n,\upsilon n)$, is systematically selected from the original data set, $\boldsymbol{x}(n,m)$. All correlation coefficients (r_a) between all pairs of the variables in the subset are calculated.

(4) The value of each of r_as is compared with the value of r_{int} specified above. If the value of any a r_a is larger than the r_{int}, then return to step (3) to continue the selection of the next subset.

(5) If all values of r_as are not larger than the r_{int}, the multiple linear regression (MLR) is then used to build a relationship model between the independent variable subset, $\boldsymbol{x}(n,\upsilon n)$, and the dependent variable set, $\boldsymbol{y}(n)$. At the same time, the relevant statistics such as the correlation coefficient (r_m) in the modeling are calculated. If the r_m is smaller than the r_{cri} determined in step (2), then return to step (3) to continue the selection of the next subset.

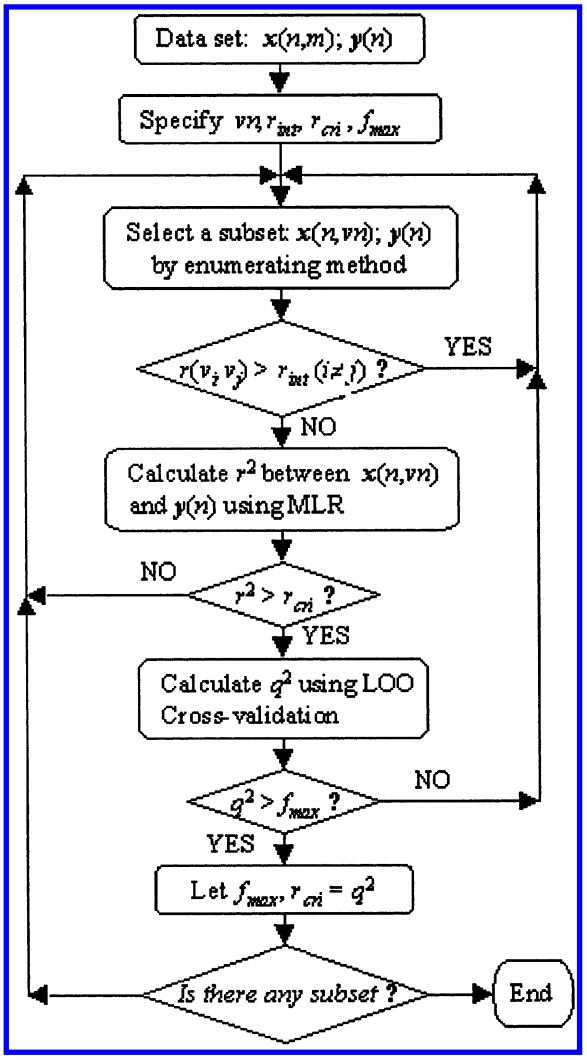

Figure 1. Selection of the optimal subset for a given υn.

(6) If the r_m is larger than the r_{cri}, the LOO cross-validation algorithm is then carried out to calculate the predictive correlation coefficient (q^2) which is compared with the f_{max} determined in the former loop. If $q^2 \leq f_{max}$, then return to step (3) to select a new subset.

(7) If $q^2 > f_{max}$, then let both f_{max} and r_{cri} equal q^2. If there still exists as any a subset not to be selected in the whole subset space, then return to step (3) to continue the selection of the next subset or the optimal process is ended.

Determination of the Best Subset. From various optimal subsets of different υn (2, 3, 4, ..., $n/5$), the best subset can be determined. It has been known that a good QSAR model should possess not only a high calibration statistics for the internal molecules but also a high predictive ability for the external molecules. It is found that the r^2 monotonically increases for increasing υn, while the q^2 gradually increases until a limited value and then decreases for increasing υn. For the root-mean-square errors (RMS), the similar results can be acquired. With the increase of υn, RMSEE is monotonically decreasing, while RMSEP gradually decreases until a limit value and then increases. So, the determination

Table 2. VSMP for the Selection of Some Optimal Subsets from among the Selwood Data Set

vn	q^2	q	RMSEP	r^2	r	RMSEE	optimal subset					
2	0.5227	0.7230	0.563	0.5985	0.7736	0.515	x_{40}	x_{50}				
3	0.6149	0.7842	0.507	0.6865	0.8286	0.455	x_4	x_5	x_{11}			
4	0.6544	0.8089	0.486	0.7769	0.8814	0.384	x_4	x_{17}	x_{40}	x_{50}		
5	0.7035	0.8387	0.445	0.7909	0.8893	0.372	x_{13}	x_{14}	x_{38}	x_{50}	x_{52}	
6	0.7184	0.8476	0.436	0.8054	0.8974	0.359	x_4	x_{13}	x_{14}	x_{38}	x_{50}	x_{52}

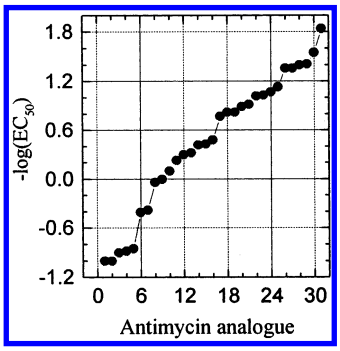

Figure 2. Distribution of activities of analogues.

of the best subset is dominated by the q or RMSEP. In this paper, the plot of RMSEP of different vn versus vn is employed together with some statistic analysis to determine the best subset entering into the final QSAR model.

RESULTS AND DISCUSSION

Data Set. The Selwood data set examined by Selwood[25] and Waller[2] is selected to test our VSMP method. This particular data set was developed as a result of a research project aimed at the development/discovery of novel antifilarials.[25] Waller employed the data set to test the novel variable selection algorithm, FRED (fast random elimination of descriptors) algorithm. He proposed that the Selwood data set can be served as a benchmark data set which can be used to evaluate the variable selection algorithms. The Selwood data set includes 31 compounds where each has the values of 53 descriptors and an in vitro biological activity, -log(EC_{50}) (EC_{50} unit: μM). In EC_{50} assay, the macrofilarial viability was determined after 120-h exposure to a range of drug concentrations, descending from 10 μM, in Eagles minimal essential medium (no serum). The values of the in vitro activities, -log(EC_{50}), are widespread and homogeneous (Figure 2). From Figure 2, eight compounds display pEC_{50} values between -1.0 and 0.0 (low activity), 12 display pEC_{50} values between 0.0 and 1.0 (moderate activity), and 10 show pEC_{50} values between 1.0 and 2.0 (high activity). The data set was rearranged by Waller.[2] The 31 compounds together with their numerical values of -log(EC_{50}) and 53 descriptors are listed in Table 1 (Supporting Information).

Optimal Subsets. The variable selection technique has been at all times combined with the model building method, and the quality of the model developed using the optimal variables acts as a criterion to evaluate the variable selection techniques. For a data set including n compounds, the maximal number of the variables among all subsets from the original data set is in general little than $n/5$. The VSMP method developed in this paper is used to search the Selwood data set which includes 31 biological activities (pEC_{50}) (dependent variable matrix, labeled as $\boldsymbol{y}(n \times 1)$) and 31×53 molecular descriptors (independent variable matrix, labeled as $\boldsymbol{x}(n \times m)$). When $vn = 2$, 3, 4, 5, and 6, respectively, five optimal subsets from the matrix, $\boldsymbol{x}(n \times m)$, selected and several corresponding statistics calculated by VSMP program with the values of $r_{int} = 0.76$, $r_{cri} = 0.20$, and $f_{max} = 0$ are listed in Table 2 where the $\boldsymbol{x}_i$ is the ith descriptor from the original Selwood data set including 53 descriptors whose names and definitions are listed in Table 3 (Supporting Information).

The Best Subset. To determine the best variable subset, the plot of the RMSEP and RMSEE versus vn is shown in Figure 3. From Figure 3 together with Table 2, although the value of q^2 is the highest when $vn = 6$, the increment of q^2 is very small from $vn = 5$ to 6, which implies that the optimal subset with five variables might be better than one with six variables. To validate this, a simple comparison of the five-variable model (eq 1) with the six-variable model (eq 2) is made. It is found that there are no significant differences between some statistics such as the RMSEE, r, F statistic, RMSEP, and q. But the difference between the regression coefficient (1.0878) and its standard deviation (0.8134) of the fourth descriptor ($\boldsymbol{x}_4$) in eq 2 is very small, which shows that the descriptor ($\boldsymbol{x}_4$) is not significant statistically. So, the best variable subset should be a combination of five descriptors of nos. $\boldsymbol{x}_{13}$, $\boldsymbol{x}_{14}$, $\boldsymbol{x}_{38}$, $\boldsymbol{x}_{50}$, and $\boldsymbol{x}_{52}$ (descriptor DIPV_Z, DIPMOM, MOFI_Y, LOGP, and SUM_F).

$$pEC_{50} = -(2.4847 \pm 0.3714) - (0.15421 \pm 0.06406)\boldsymbol{x}_{13} - (0.098986 \pm 0.038679)\boldsymbol{x}_{14} - (6.0448 \pm 1.1810) \times 10^{-5}\boldsymbol{x}_{38} + (0.55902 \pm 0.08253)\boldsymbol{x}_{50} + (1.7913 \pm 0.4178)\boldsymbol{x}_{52} \quad (1)$$

$$n = 31,\ m = 5,\ r^2 = 0.7909,\ r = 0.8893,\ \text{RMSEE} = 0.372,\ F = 18.914 \text{ (modeling)}$$

$$q^2 = 0.7035,\ q = 0.8387,\ \text{RMSEP} = 0.445 \text{ (LOO prediction)}$$

$$pEC_{50} = -(2.2636 \pm 0.4013) + (1.0878 \pm 0.8134)\boldsymbol{x}_4 - (0.15111 \pm 0.06312)\boldsymbol{x}_{13} - (0.093340 \pm 0.038316)\boldsymbol{x}_{14} - (6.0669 \pm 1.1629) \times 10^{-5}\boldsymbol{x}_{38} + (0.54943 \pm 0.08158)\boldsymbol{x}_{50} + (1.8033 \pm 0.4115)\boldsymbol{x}_{52} \quad (2)$$

$$n = 31,\ m = 6,\ r^2 = 0.8054,\ r = 0.8974,\ \text{RMSEE} = 0.359,\ F = 16.556 \text{ (modeling)}$$

$$q^2 = 0.7184,\ q = 0.8476,\ \text{RMSEP} = 0.436 \text{ (LOO prediction)}$$

Comparison with FRED Method. Waller[2] proposed a novel variable selection technique, FRED, to search the best subset from the benchmark data set and developed a six-variable QSAR model with the q^2 value of 0.6952 and corresponding r^2 value of 0.8288. The six variables are

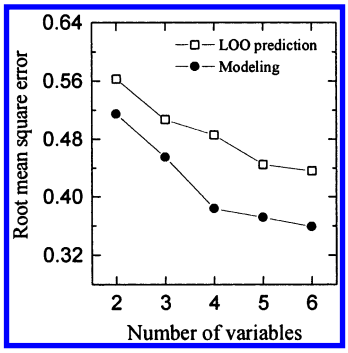

Figure 3. RMSEE and RMSEP vary with vn.

ATCH_4 (x_4), ESDL_3 (x_{17}), VDWVOL (x_{35}), MOFI_X (x_{37}), LOGP (x_{50}), and SUM_F (x_{52}). The QSAR model is expressed as follows.

$$pEC_{50} = (2.0343 \pm 1.0302) + (3.0868 \pm 0.9402)x_4 + (0.61250 \pm 0.26845)x_{17} - (0.012861 \pm 0.003009)x_{35} + (4.31 \pm 1.07) \times 10^{-4} x_{37} + (0.42096 \pm 0.07875)x_{50} + (0.46253 \pm 0.47418)x_{52} \quad (3)$$

$$n = 31,\ m = 6,\ r^2 = 0.8288,\ r = 0.9104,\ \text{RMSEE} = 0.336,\ F = 19.359\ \text{(modeling)}$$

$$q^2 = 0.6952,\ q = 0.8338,\ \text{RMSEP} = 0.457\ \text{(LOO prediction)}$$

Apparently, the significance statistically of the variable x_{52} is very poor although the model has high q^2 and r^2. Eliminating the x_{52}, rebuild a five-variable linear model (eq 4) using MLR technique. From eq 4, although the calibrating quality of the model ($r^2 = 0.8220$) is slightly poorer than one of eq 3 ($r^2 = 0.8288$), both the predictive quality ($q^2 = 0.6955$) and statistically significance ($F = 23.085$) are better than eq 3 ($q^2 = 0.6952$ and $F = 19.359$).

$$pEC_{50} = (2.6044 \pm 0.8476) + (3.1420 \pm 0.9376)x_4 + (0.76317 \pm 0.21935)x_{17} - (0.013698 \pm 0.002881)x_{35} + (4.35 \pm 1.06) \times 10^{-4} x_{37} + (0.44232 \pm 0.07558)x_{50} \quad (4)$$

$$n = 31,\ m = 5,\ r^2 = 0.8220,\ r = 0.9066,\ \text{RMSEE} = 0.343,\ F = 23.085\ \text{(modeling)}$$

$$q^2 = 0.6955,\ q = 0.8340,\ \text{RMSEP} = 0.457\ \text{(LOO prediction)}$$

Comparison of our VSMP results (eqs 1 and 2) with Waller's FRED results (eqs 3 and 4) reveals that a model with higher r^2 has not always a higher q^2 though r^2 is always higher than q^2 for a given vn. We consider that a good QSAR model should have not only enough high estimating correlation coefficient (r^2) and LOO predictive correlation coefficient (q^2) but also significance statistically for all variables entering the model. Golbraikh and Tropsha[23] also indicated that a high q^2 is the necessary condition of a model with high predictive power but not the sufficient condition. So, variable selection technique is above all dependent on the predictive process. Because LOO cross-validation is one of the simplest cross-validation techniques and easily performed, in our VSMP method, the q^2 obtained in LOO procedure is used to control the search of various optimal subsets.

To further explain that our VSMP method can find out the best subset which makes the QSAR based on the subset have the best predictive power, a 26-variable subset obtained by six different search algorithms (see Table 3) summarized in the literature[23] from among the original data set including 53 descriptors is also used to do VSMP analysis. The 26 descriptors are x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_{11}, x_{12}, x_{13}, x_{17}, x_{19}, x_{22}, x_{24}, x_{26}, x_{35}, x_{36}, x_{37}, x_{38}, x_{39}, x_{40}, x_{41}, x_{50}, x_{51}, x_{52}, and x_{53}, respectively. On the other hand, to compare with the other algorithms based on the calibration or estimation process, the VSMP program is rerun by replacing the $r_{cri} = q^2$ with the $r_{cri} = r^2$. The relevant results are listed in Table 4 where the numerical values in parentheses refer to the number of the descriptors used to do VSMP analysis. From Table 4, for five-variable linear model, both the r^2−based VSMP and FRED give the same results, but the values of q^2 are lower than one of the q^2−based VSMP. For the six-variable model, the r^2−based VSMP presents the highest r^2 (0.8369) and high q^2 (0.7171), which shows that the FRED method does not find out the best subset based on the estimation when $vn = 6$. It is found that the variable subset optimized by FRED when the x_{52} replaced by the x_{13} is the subset obtained by the r^2−based VSMP. The interrelation coefficient of 0.1139 between the x_{52} and x_{13} also explains that the VSMP method has more chances to find out the best subset. However, if LOO q^2 is considered as a criterion, the q^2−based VSMP gives the highest value of q^2 (0.7035 and 0.7184) both for five- and six-variable models.

Varieties of q^2 and r^2 with vn. To determine the best subset from various optimal subsets of different vn, it is essential to study the varieties of q^2 and r^2 with increasing vn. The 26 descriptors from the literature[2] are used to examine the varieties of q^2 and r^2 with increasing vn. In this literature, various descriptors (with the symbol "√" in Table 3 (Supporting Information)) optimized by six different algorithms, i.e., FRED, MUSEUM (mutation and selection uncover models), GFA (genetic function approximation), CSA (cluster significance analysis), NN (neural network), and NLM/SR (nonlinear mapping/stepwise regression), were listed in Table 3 (Supporting Information). From Table 3 (Supporting Information), the number of descriptors in the optimal subset selected by FRED, MUSEUM, GFA, CSA, NN, or NLM/SR techniques is $m = 13$, 15, 10, 7, 8, or 9, respectively. Our VSMP approach is still employed to search these optimal descriptor sets. The results shown in Figure 4 illustrate the varied curves of RMSEP and RMSEE with increasing vn. From Figure 4, the following conclusions can be deduced.

(1) The root-mean-square error in modeling step (RMSEE) is always descended for increasing number of variables in an optimal subset.

(2) RMSEE is always little than RMSEP.

(3) The variety of the root-mean-square error in LOO prediction step (RMSEP) with vn seems to exist as a parabola-like mode. That is, the RMSEP is primary descended until a relative low RMSEP value and then risen slowly for the increasing vn. As shown in Figure 4, in most cases, the parabola is complete (such as in FRED, MUSEUM, GFA, CSA, and NN), which states that there exists

Table 4. Result Comparison of q^2− and r^2−Based VSMP to FRED

method	descriptor	q^2	RMSEP	r^2	RMSEE	F
FRED	x_4, x_{17}, x_{35}, x_{37}, x_{50}	0.6955	0.457	0.8220	0.343	23.085
FRED	x_4, x_{17}, x_{35}, x_{37}, x_{50}, x_{52}	0.6952	0.457	0.8288	0.336	19.359
q^2-based VSMP (53)	x_{13}, x_{14}, x_{38}, x_{50}, x_{52}	0.7035	0.445	0.7909	0.372	17.914
q^2-based VSMP (53)	x_4, x_{13}, x_{14}, x_{38}, x_{50}, x_{52}	0.7184	0.436	0.8054	0.359	16.556
q^2-based VSMP (26)	x_{13}, x_{14}, x_{38}, x_{50}, x_{52}	0.7035	0.445	0.7909	0.372	17.914
q^2-based VSMP (26)	x_4, x_{13}, x_{14}, x_{38}, x_{50}, x_{52}	0.7184	0.436	0.8054	0.359	16.556
r^2-based VSMP (53)	x_4, x_{17}, x_{35}, x_{37}, x_{50}	0.6955	0.457	0.8220	0.343	23.085
r^2-based VSMP (53)	x_4, x_{13}, x_{17} ,x_{35}, x_{37}, x_{50}	0.7171	0.442	0.8369	0.328	20.520
r^2-based VSMP (26)	x_4, x_{17}, x_{35}, x_{37}, x_{50}	0.6955	0.457	0.8220	0.343	23.085
r^2-based VSMP (26)	x_4, x_{13}, x_{17} ,x_{35}, x_{37}, x_{50}	0.7171	0.442	0.8369	0.328	20.520

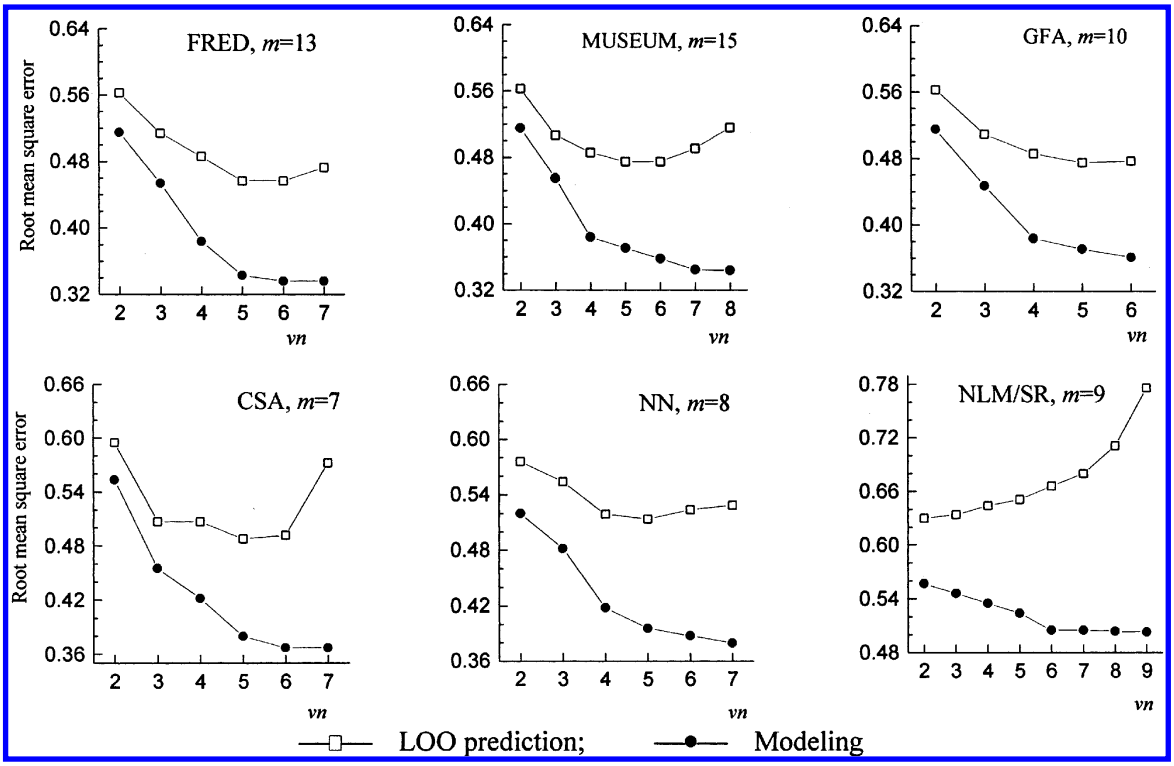

Figure 4. Various varieties of RMSEE and RMSEP with vn.

as the best subset from the original data set, $\mathbf{x}(n,m)$. In the other conditions, the plot of RMSEP versus vn can be just a part of the parabola. For example, the curve obtained by our VSMP analysis on the nine descriptors optimized by the NLM/SR algorithm is just the latter half of the parabola, while the RMSEP curve shown in Figure 3 is just the former half. Besides, it is also found that the former section of the parabola is often displayed when the number of the original descriptors (m) is large. If the VSMP analysis is performed on all 26 variables ($m = 26$) employed by the above six algorithms, the varieties of RMSEE and RMSEP with increasing vn are shown in Figure 5, which is similar to the RMSEP curve in Figure 3. In such a case, the determination of the best subset needs to examine sufficiently the q^2, r^2, the significance of regression coefficient, and the allowable interrelation coefficient between the variables.

The conclusions above further explain why the RMSEP or q^2 in LOO prediction step is more important than the RMSEE or r^2 in modeling process for variable selection and QSAR models. This is because the variety of RMSEP with vn disagrees with one of the RMSEE and a model with high r^2 or low RMSEE does not always have high q^2 or low RMSEP.

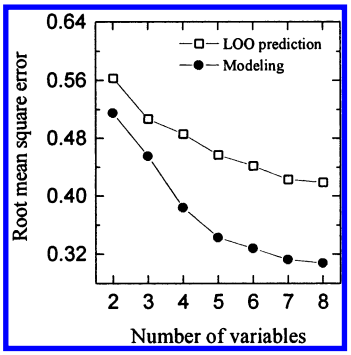

Figure 5. Plot of RMS versus vn for 26 variable set.

Two Important Controlled Factors. The VSMP developed in our laboratory is a revised all possible subset (ASR) or best subset regression technique. Here two important controlled parameters, the interrelation coefficient between the variables (r_{int}) and predictive correlation coefficient obtained in LOO cross-validation process (q^2), were introduced into the classical ASR algorithm to search the optimal variable subsets for different number of the variables. The q^2 is used to ensure a high enough predictive power of the model in the condition of high r^2. At the same way, the q^2

together with the r_{int} control the running speed of the VSMP program. The r_{int} will make the MLR calculation of the subset with one or more higher interrelation coefficients than r_{int} not perform. The q^2 or more exactly f_{max} controls whether the LOO cross-validation calculation after MLR process is performed or not. It is just two factors that improve the performance of classical ASR algorithm, the ASR running speed and translation of the calculation based on modeling into the calculation based on prediction.

CONCLUSION

A novel variable selection and modeling technique based on the prediction, VSMP, is developed. The technique improves the performance of classical all-subsets regression (ASR) by the introduction of two parameters, the interrelation coefficient between the variables (r_{int}) and predictive correlation coefficient in LOO cross-validation process (q^2). It differs from the ASR due to its higher running speed and both high calibrated statistics for the internal molecules and predictive power for the external molecules. The validity of the VSMP has been tested using a benchmark data set, the Selwood data set consisting of 31 antifilarial antimycin analogues where each has the value of 1 in vitro activity and 53 descriptors. It can be foreseen that the method will become one of general variable selection and QSAR research tools based on the predictive process. Related studies are in progress.

ACKNOWLEDGMENT

We are especially grateful to the China Postdoctoral Science Foundation and the National High Technology Research and Development Program of China (No. 2001AA646010) for their financial support.

Supporting Information Available: Tables of biological activity and parameter values of 31 compounds in the Selwood data set (Table 1) and names and descriptions of 53 descriptors and descriptors selected primarily by six algorithms (Table 3). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Yasri, A.; Hartsough, D. Toward an optimal procedure for variable selection and QSAR model building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218−1227.
(2) Waller, C. L.; Bradley, M. P. Development and validation of a novel variable selection technique with application to multidimensional quantitative structure−activity relationship studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 345−355.
(3) Steyerberg, E. W.; Eijkemans, M. J. C.; Habbema, J. D. F. Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. *J. Clin. Epidemiol.* **1999**, *52*(10), 935−942.
(4) Agostinelli, C. Robust stepwise regression. *J. Appl. Stat*. **2002**, *29*(6), 825−840.
(5) Westfall, P. H.; Young, S. S.; Lin, D. K. J. Forward selection error control in the analysis of supersaturated design. *Stat. Sinica* **1998**, *8*, 101−117.
(6) Hoskuldsson, Agnar. Variable and subset selection in PLS regression. *Chemom. Intell. Lab. Syst*. **2001**, 55, 23−38
(7) Geladi, P.; Kowalski, B. R. Partial least-sqaures regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(8) Norinder, U.; Rivera, C.; Unden, A. A quantitative structure−activity relationship study of some substance P-related peptides. A multivariate approach using PLS and variable selection. *J. Pept. Res*. **1997**, *49*(2), 155−162.
(9) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc*. **1988**, *110*, 5959−5967.
(10) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure−Activity Studies. *J. Med. Chem*. **1999**, *42*, 573−583.
(11) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of quantitative structure−activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 669−677.
(12) Kubinyi, H. Variable selection in QSAR studies, 1. An evolutionary algorithm. *Quant. Struct.−Act. Relat.* **1994**, *13*, 285−294.
(13) Wikel, J.; Dow, E. The use of neural networks for variable selection in QSAR. *Bioorg. Med. Chem. Lett*. **1993**, *3*, 645−651.
(14) Kovalishyn, V. V.; Tetko, I. V.; Luik, A. I.; Kholodovych, V. V.; Villa, A. E. P.; Livingstone, D. J. Neural network studies. 3. Variable selection in the cascade-correlation learning architecture. *J. Chem. Inf. Comput. Sci*. **1998,** *38*(4), 651−659.
(15) Ramadan, Z.; Song, X. H.; Hopke, P. K.; Johnson, M. J.; Scow, K. M. Variable selection in classification of environmental soil samples for partial least-squares and neural network models. *Anal. Chim. Acta* **2001,** *446*(1−2), 233−244.
(16) Goldberg, D. E. *Genetic algorithms in search, optimization and machine learning*; Addison-Wesley: New York, 1989.
(17) Hasegawa, K.; Kimura, T.; Funatsu, K. GA strategy for variable selection in QSAR studies: Application of GA-based region selection to a 3D-QSAR study of acetylcholinesterase inhibitors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*(1), 112−120.
(18) 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*, 4, 854−866.
(19) Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom*. **1992**, *6*, 267−281.
(20) Abraham, B.; Chipman, H.; Vijayan, K. Some risks in the construction and analysis of supersaturated design. *Technometrics* **1999**, *41*, 135−141.
(21) Smith, J. S.; Macina, O. T.; Sussman, N. B.; Luster, M. I.; Karol, M. H. A robust structure−activity relationship (SAR) model for esters that cause skin irritation in humans. *Toxicol. Sci*. **2000**, *55*(1), 215−222.
(22) Liu, S. P.; Lu, J. C.; Kolpin, D. W.; Meeker, W. Q. Analysis of environmental data with censored observations. *Environ. Sci. Technol.* **1997**, *31*(12), 3358−3362.
(23) Golbraikh, A.; Tropsha, A. Beware of *q* 2. *J. Mol. Graph. Model*. **2002**, *20*, 269−276.
(24) Liu, S. S.; Yin, C. S.; Wang, L. S. Combined MEDV-GA-MLR Method for QSAR of Three Panels of Steroids, Dipeptides, and COX-2 Inhibitors. *J. Chem. Inf. Comput. Sci*. **2002**, *42*(3), 749−756.
(25) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure−activity relationships of antifilarial antimycin analogues: a multivariate pattern recognition study. *J. Med. Chem*. **1991**, *33*, 136−142.

CI020377J