Optimized Block-wise Variable Combination by Particle Swarm Optimization for Partial Least Squares Modeling in Quantitative Structure−Activity Relationship Studies

Wei-Qi Lin, Jian-Hui Jiang, Qi Shen, Guo-Li Shen, and Ru-Qin Yu*

State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China

Received March 30, 2004

The use of numerous descriptors that are indicative of molecular structure is becoming common in quantitative structure−activity relationship (QSAR) studies. As all of the descriptors might carry more or less molecular information, it seems more advisable to investigate the possible variable combination rather than variable selection. In this paper, an optimized block-wise variable combination (OBVC) by particle swarm optimization based on partial least squares modeling has been proposed for variable combination. An *F* statistic is also introduced to determine the dimensionality of the PLS model. The performance is assessed using two QSAR data sets. Experimental results have shown the good performance of this technique compared to those obtained by stepwise regression.

1. INTRODUCTION

Quantitative structure−activity relationships (QSARs) relating chemical structure to biological activity is an important area of chem-bioinformatics. There are many descriptors representing chemical structure such as spatial, electronic, topological, thermodynamic, quantum mechanical, and shape descriptors in QSARs studies. The number of compounds with the biological activity values available is usually small compared with the number of structural descriptors. It is well-known that too many variables could lead to possible overfitting, and too few variables would cause underfitting of the model resulting in a low correlation between structure and activities (QSAR) or properties (QSPR).

Selecting the descriptors that are really indicative of the biological activity concerned becomes one of the key steps in QSAR studies. There have been many variable selection methods suggested, such as simulated annealing,[1] stepwise regression,[2−4] evolutionary algorithms,[5,6] genetic algorithms,[7,8] and so on. However, it remains a difficult task as there are no generally accepted rules to guide this selection. It is dangerous to remove any variable for it might carry more or less information related to the activities of the molecules and might be helpful to improve the regression against the property even if it is less informative. It seems that one had better investigate the possible combinations of variables for finding the best model. However, few works were done to address this problem. Very recently Du and co-workers[9] proposed a procedure to use a subjective criterion for allocating the variables into several blocks followed by a single representative variable extraction based on canonical correlation analysis (CCA).

In this paper, an optimized block-wise variable combination (OBVC) by particle swarm optimization for partial least squares modeling has been proposed to carry out the optimized combinations of variables, from which one extracts the most related latent variables that capture maximally the information of the original variable blocks to establish the regression model. Particle swarm optimization (PSO) algorithm was modified to seek the optimized variable combination. Partial least squares (PLS), an efficient technique to reduce the data dimensionality, was used for building QSAR models. And an *F* statistic was introduced to make the procedure of the variable extraction proceed automatically without interference of the user. PSO developed by Eberhart and Kennedy[10−13] in 1995 is a stochastic global optimization technique inspired by social behavior of bird flocking. The algorithm models the exploration of a problem space by a population of individuals or particles. Similar to GAs and EAs, PSO is a population based optimization tool, which searches for optima by updating generations. However, each individual in PSO flies in the search space with a velocity that directs the flying of the particle instead of crossover and mutation operators. Compared to GAs and EAs, the advantages of PSO are that it is easy to implement and there are few parameters to adjust. Using the modified PSO, the optimized variable combination can be obtained objectively. Thus, the variables belonging to the same block might be considered as an ensemble named variable block, which includes all individual variables in this block. Within each block, most related latent variables, which possess the maximum correlation with property, were extracted with the help of PLS, and an *F* statistic was used to decide the dimensionality of the PLS model. Then, the structure−activity correlation model including these new variables is established. The regression model achieved in this way shows significant improvement both in fitting and prediction ability.

This method was used to predict carcinogenic potency of aromatic amines and antagonism of angiotensis II antagonists. Aromatic amines are widely used in papermaking, leather, textile, and other industries. Most kinds of aromatic amines are mutagenic and carcinogenic. The vasoactive hormone angiotensin II produced by the renin-angiotensis system plays

* Corresponding author phone: +86-731-8822577; fax: 86 731 8822782; e-mail: rqyu@hnu.net.cn.

10.1021/ci049890i CCC: $30.25
Published on Web 01/20/2005

an integral role in the pathophysiology of hypertension. Angiotensin-converting enzyme (ACE) inhibitors, which block the converting of angiotensin I to the angiotensin II, are widely used for the treatment of hypertension and congestive heart failure. The results compared to those obtained by the classical stepwise regression show the good performance of the technique. It has been demonstrated that optimized block-wise variable combination by particle swarm optimization for partial least squares modeling is effective in QSAR studies and can be used as a complementary tool, for the experimental assessment might be expensive, hazardous, and time-consuming.

2. THEORY

2.1. Modified Particle Swarm Optimization. Particle swarm optimization (PSO) involves simulating social behavior among individuals (particles) "flying" through a multidimensional search space, each particle keeps track of its coordinates in the problem space which are associated with the best solution (fitness) it has achieved so far. The first step of the algorithm is to randomly initialize the position and velocity of each particle in the swarm, dispersing them uniformly across the search space. The *i*th particle is represented as $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iD})$. Velocity, the rate of the position change for particle *i* is represented as $\mathbf{v}_i = (v_{i1}, v_{i2}, ..., v_{iD})$. PSO postulates that particles should move toward some combination of their personal best position and the global best position. The personal best position $\mathbf{p}_i = (p_{i1}, p_{i2}, ..., p_{iD})$ is the best previous position of the *i*th particle that gives the best fitness value. The global best position $\mathbf{p} = (p_{g1}, p_{g2}, ..., p_{gD})$ is the best particle among all the particles in the population. In every iteration, each particle is updated by following the two best values.

Most versions of PSO have operated in continuous and real-number space. For a discrete problem expressed in an integer string, which varies from 1 to *k*, a particle moves in a search space restricted to integer from 1 to *k* on each dimension. Here *k* is the number of the variable blocks. In an integer string problem, updating a particle represents changes of a site that should be an integer, and the velocity represents the probability of site x_{id} taking the integer value from 1 to *k* respectively.

In the PSO algorithm, a population of particles is updated on the basis of information about each particle's previous best performance and the best particle in the population. According to an information sharing mechanism of PSO, a modified discrete PSO[14] was proposed as follows. The velocity v_{id} of every individual is a random number in the range of (0,1). The resulting change in position is then defined by the following rule

$$\text{If } (0 < v_{id} \le a), \text{ then } x_{id}(\text{new}) = x_{id}(\text{old}) \quad (1)$$

$$\text{If } (a < v_{id} \le (1 + a)/2), \text{ then } x_{id}(\text{new}) = p_{id} \quad (2)$$

$$\text{If } ((1 + a)/2 < v_{id} \le 1), \text{ then } x_{id}(\text{new}) = p_{gd} \quad (3)$$

where *a* is a random value in the range of (0,1) named static probability. Static probability *a* started with a value of 0.5 and decreases to 0.33 when the iteration terminates. Though the velocity in the modified discrete PSO is different from that in the continuous version of PSO, an information sharing mechanism and updating model of particle by following the two best positions is the same in the two PSO versions.

To circumvent convergence to local optima and improve the ability of the modified PSO algorithm to overcome local optima, five percent of particles are randomly selected, and each site of the selected particles has a probability of 0.5 to vary the value in a stochastic manner. If the minimum error criterion is attained or the number of cycles reaches a user-defined limit, the algorithm is terminated.

Using decreasing static probability and some percent of randomly fling particles to overcome local optima, the modified PSO remains having satisfactory converging characteristics.

2.2. Optimized Block-wise Variable Combination (OBVC) Guided by PSO. In QSARs studies, there are many descriptors representing chemical structures such as spatial, electronic, topological, thermodynamic, quantum mechanical, and shape descriptors. One may notice that many descriptors are similar. However, it is still unadvisable to remove any variable which seems less informative, as the variable might be helpful to improve the regression against the property. An apt solution is allocating the variable into the *k* blocks stochastically first and then searching the optimized block-wise variable combination by PSO followed by the most related latent variable extraction based on partial least squares modeling. These latent variables obtained are used to substitute the original variables for QSAR model building. Here *k* increases one by one and is decided ultimately according to the curve of the minimum residual sum of the squares of the model vs *k*. The original data set consisting of an independent variable matrix, $\boldsymbol{x}(i,j)$, which includes *J* descriptors of *I* compounds, and a dependent variable matrix, $\mathbf{y}(i)$, including the properties (biological activities or physiochemical properties) of the compounds, is autoscaled according to the following rule:

$$x_{ij,\text{new}} = \frac{x_{ij,\text{old}} - m_j}{C_j} \quad (4)$$

$$y_{i,\text{new}} = y_{i,\text{old}} - m_i \quad (5)$$

In this expression, C_j is the variance of the variable *j*, m_j is the mean of the variable *j*, and m_i is the mean of the variable $\mathbf{y}(i)$. *J* descriptors are grouped as *k* blocks stochastically to get variable blocks $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_k$. Within each block, the most related latent variables, which capture as fully as possible the information of the original variable block, are extracted by PLS.

The dimensionality of the PLS model in each block is decided by an *F* statistic which is introduced into this procedure to select latent variables. The *F* statistic is defined by the following rule:

$$F = \frac{\frac{1}{I-n}(\text{RSS}_{n-1} - \text{RSS}_n)}{\frac{1}{I-n-1}\text{RSS}_n} \quad (6)$$

Here *I* is the number of the compounds and *n* is the number of the latent variables extracted from each block. When *F* is less than or equal to 1, that is to say the newly extracted latent variable does not reduce the error of the models markedly, the latent variable will be abnegated and the

procedure of extracting the latent variables from the block will stop. Otherwise the procedure will go on. Then the latent variables extracted from the blocks, which are very few compared to the original data set, can be used to substitute the original variables to establish PLS model against the property **y**. In this way, the model established between descriptors and the property will be simplified with fewer variables but without losing correlation information. Optimized block-wise variable combination (OBVC) by particle swarm optimization for partial least squares modeling is carried out orderly in the following steps.

Step1. Randomly initialize all the initial integer strings in modified discrete PSO with an appropriate size of population. The integer strings are restricted to integer from 1 to *k*.

Step 2. Calculate the fitness function of an individual corresponding to models in the training set. If the best objective function of the generation fulfills the end condition, the training is stopped with the results output, otherwise, go to the next step.

Step 3. Update the population according to the modified discrete PSO.

Step 4. Go back to the second step to calculate the fitness of the renewed population.

2.3. QSAR Model Generated by OBVC Algorithm. Suppose variable blocks $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_K$ include j_1, j_2, ..., j_K descriptors, respectively.

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K) \tag{7}$$

$$\sum_{k=1}^{K} j_k = J \tag{8}$$

$$\mathbf{Z}_k = \mathbf{X}_k \mathbf{V}_k \quad k = 1, 2, ..., K \tag{9}$$

V is the combination weight of the descriptors and can be written as follows[15]

$$\mathbf{V} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1} \tag{10}$$

where **W** is the loading weight of **X**, and **P** is the estimated loading of **X**. Then,

$$\mathbf{Z} = (\mathbf{Z_1}, \mathbf{Z_2}, ..., \mathbf{Z_K}) = \mathbf{X}\begin{pmatrix} V_1 & & & \\ & V_2 & & \\ & & \ddots & \\ & & & V_K \end{pmatrix} \tag{11}$$

where

$$\begin{pmatrix} V_1 & & & \\ & V_2 & & \\ & & \ddots & \\ & & & V_K \end{pmatrix}$$

is a quasi-diagonal matrix with quasi-diagonal entries being the matrices V_1, V_2, ..., V_K.

The calibration equation in traditional univariate calibration is simply the linear predictor

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{12}$$

Then,

$$\mathbf{Y} = \mathbf{ZR} = \mathbf{X}\begin{pmatrix} V_1 & & & \\ & V_2 & & \\ & & \ddots & \\ & & & V_K \end{pmatrix}\mathbf{R} \tag{13}$$

$$\mathbf{R} = \mathbf{W}(\mathbf{P}^{\mathbf{T}}\mathbf{W})^{-1}\mathbf{Q} \tag{14}$$

In this expression, **Q** is the loadings of **Y**. Accordingly the regression coefficient **B** can be obtained by the following formula:

$$\mathbf{B} = \begin{pmatrix} V_1 & & & \\ & V_2 & & \\ & & \ddots & \\ & & & V_k \end{pmatrix}\mathbf{R} \tag{15}$$

It is noted that the deduction of the exact model from eqs 7 to 15 might sometime be useful for the interpretation of the QSAR model. In general settings, a small combination weight (absolute value) **W** implies the corresponding descriptor may be closely correlated to the other descriptors in the block or it is of little importance to the model, and an increased absolute value of **B** means the model depends more strongly on the corresponding descriptor.

3. DATA SETS

3.1. Aromatic Amine Data. A set of 41 aromatic amines, whose carcinogenic potencies are collected together in a comprehensive review by Benigni et al.,[16] was used to test the performance of the algorithm in the present paper. The data set was stochastically divided into two groups. Thirty-one compounds were used as the training set for developing regression models, while the remaining 10 compounds were used as the validation set in the prediction of carcinogenic potency of aromatic amines. The carcinogenic potency is expressed as $\mathrm{BRR} = \log(\mathrm{MW}/\mathrm{TD}_{50})_{\mathrm{rat}}$, where MW is the molecular weight, and TD_{50} is the daily dose rate necessary to halve the probability of an experimental rat remaining tumorless to the end of its standard life span. A series of molecular descriptors was calculated for the amines as the original variables. These include topological thermodynamic, electronic, spatial, and information-content descriptors. The topological descriptors calculated include electrotopological-state indices (E-state indices)[17−21] representing the electron accessibility such as S-ss$\mathrm{CH_2}$, S-aaCH,[21] etc. The E-state index for an atom represents the electron accessibility associated with each atom type. It is an indication of the presence/absence of a given atom type and the count of the number of atoms of a particular element type. For example, in the symbol S-ss$\mathrm{CH_2}$, 's' stands for the sum of E-state values for all the $-\mathrm{CH_2}-$ groups in the molecule, 'ss' stands for the two single bonds of that group, and '$\mathrm{CH_2}$' represents the formula of the hydride group. The thermodynamic descriptors were taken describing the hydrophobic character (logP, logarithm of the partition coefficient in octanol/water),[22] refractivity (MolRef, molar refractivity)[22] and thermal stability of the molecules (Hf, heat of formation),[23] and the dissolution free energy for water and octanol (Fh2o, desolvation free energy for $\mathrm{H_2O}$; Foct, desolvation free energy for octanol).[24] The electronic descriptors taken were

Table 1. Descriptions of the Variables in the Data Set 1

variable	description	block	variable	description	block
1. S-aaCH	E-state index	1	40. LUMO	lowest unoccupied molecular orbital energy	1
2. S-aasC	E-state index	1	41. RadOfGyration	radius of gyration	2
3. S-sNH$_2$	E-state index	2	42. area A^2	molecular surface area	2
4. S-sCH$_3$	E-state index	1	43. MW g/mol	molecular weight	1
5. S-ssCH$_2$	E-state index	1	44. Vm A^3	molecular volume	1
6. S-ssssC	E-state index	1	45. density g/mL	density	2
7. S-dsN	E-state index	1	46. PMI-mag	principal moment of inertia	2
8. I-sCH$_3$	E-state index	1	47. PMI-X	principal moment of inertia	2
9. I-sNH$_2$	E-state index	2	48. PMI-Y	principal moment of inertia	2
10. I-ssCH$_2$	E-state index	1	49. PMI-Z	principal moment of inertia	2
11. N-sCH$_3$	E-state index	1	50. PMI	principal moment of inertia	2
12. N-aaCH	E-state index	2	51. Hbond acceptor	number of hydrogen bond acceptors	1
13. N-aasC	E-state index	1	52. Hbond donor	number of hydrogen bond donors	1
14. N-sNH$_2$	E-state index	2	53. RotBonds	number of rotatable bonds	1
15. N-dsN	E-state index	1	54. Jurs-WPSA-2	Jurs charged partial surface area descriptor	2
16. N-ssO	E-state index	2	55. Jurs-TPSA	Jurs charged partial surface area descriptor	1
17. I-dNH	E-state index	1	56. Jurs-RPSA	Jurs charged partial surface area descriptor	2
18. S-dNH	E-state index	2	57. Jurs-RASA	Jurs charged partial surface area descriptor	1
19. JX	Balaban indices	2	58. Shadow-Zlength	shadow indices	2
20. Kappa-1	Kappal topological indices	1	59. Shadow-nu	surface area projections	2
21. Kappa-3	Kappal topological indices	1	60. CIC	multigraph information content indices	2
22. SC_1	Kier & Hall subgraph count index	1	61. BIC	multigraph information content indices	2
23. CHI-1	molecular connectivity index	1	62. SIC	multigraph information content indices	1
24. logZ	topological descriptor	2	63. IC	multigraph information content indices	1
25. AlogP	total value of logP	2	64. V-ADJ-mag	information indices	1
26. LogP	the octanol/water partition coefficient	1	65. E-ADJ-mag	information indices	2
27. MR$_{CM^{**}-3}$	molar refractivity	1	66. V-DIG-mag	information indices	2
28. MolRef	molar refractivity	2	67. E-DIST-mag	information indices	1
29. Hf	heat of formation	1	68. IAC-Total	information of atomic composition index	2
30. FH2o	desolvation free energy for water	2	69. MR$_3$	molar refractivity	2
31. Foct	desolvation free energy for octanol	2	70. ΣMR$_{2,6}$	molar refractivity	2
32. Apol	sum of atomic polarizabilities	2	71. ΣR$_{2,6}$	inductive electronic substituent constants	1
33. Dipol-mag	dipole	1	72. ΣF$_{2,6}$	resonance-polar electronic substituent constants	1
34. Dipol-X	dipole	1	73. Es(R)	steric properties of substituents	2
35. Dipol-Y	dipole	2	74. I(Bi)	indicator variables	1
36. Dipol-Z	dipole	1	75. I(F)	indicator variables	1
37. Sr	superdelocalizability	2	76. I(BiBr)	indicator variables	1
38. Energy	energy of conformation	1	77. I(RNNO)	indicator variables	1
39. HOMO	highest occupied molecular orbital energy	2			

concerning surperdelocalizability (Sr), atomic polarizabilities (Apol),[25] the dipole moment (Dipole),[25] highest occupied molecular orbital energy (HOMO), and lowest unoccupied molecular orbital energy (LUMO). The spatial descriptors used involve radius of gyration (RadOfGyration), Jurs charged partial surface area descriptors (Jurs descriptors),[26] density, principal moment of inertia (PMI),[26] and molecular volume. Some so-called information-content descriptors,[27] such as atomic composition indices and multigraph information-content indices, were also included in the list. Besides the aforementioned nearly calculated molecular descriptors, 9 variables used by Benigni et al. were also included in the list of variables. *I*(Bi), *I*(BiBr), *I*(F), and *I*(RNNO) are indicator variables that characterize some special features of the compounds. (*I*(Bi) = 1 for biphenylamines; *I*(BiBr) = 1 for biphenylamines with a bridge between the phenyl rings; *I*(F) = 1 for aminofluorenes; *I*(RNNO) = 1, if the amino group is substituted with (Me)NO). Table 1 summarizes all molecular descriptors used as the original variables.

3.2. Angiotensin II Antagonists Data. A set of 38 4H-1,2,4-trialzoles as angiotensin II antagonists, whose antagiosm are collected together in a comprehensive review by Kurup et al.,[28] was used to further check the validity of the proposed methods. Ashton et al.[29] synthesized and evaluated 4H-1,2,4-triazoles for their antagonism against angiotensin II, which are expressed as IC$_{50}$, the molar concentration of the compound causing 50% antagonism of angiotensin II.

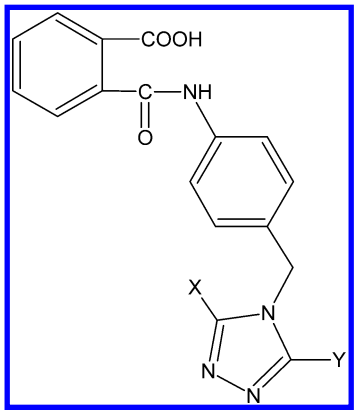

Figure 1. Structure of 4H-1,2,4-triazole.

The data set of 38 4H-1,2,4-triazoles was randomly divided into two groups with 28 compounds used as the training set for developing the regression models and the remaining 10 compounds used as the validation set in the prediction of antagonism against angiotensin II 4H-1,2,4-triazoles. Molecular structure and numbering of substituents in the series of 4H-1,2,4-triazoles are represented in Figure 1.

A series of molecular descriptors was calculated for analogues of 4H-1,2,4-triazoles including structural, spatial,

Table 2. List of Molecular Descriptors for 4H-1,2,4-Triazoles Studied as Original Variables

variable	block	variable	block
1.S_aaCH[a]	1	20. Dipol-Z[a]	3
2.S_aasc[a]	1	21. Sr[a]	2
3.S_dO[b]	2	22. HOMO[a]	1
4. S_aaN[b]	1	23. LUMO[a]	3
5. S_sCH3[a]	1	24. RadOfGyration[a]	1
6. S_ssO[b]	3	25. Area[a]	3
7. S_dssC[b]	1	26. MW[a]	2
8. S_ssS[b]	2	27. Vm[a]	2
9. S_aasN[b]	1	28. Density[a]	1
10. AlogP[a]	2	29. PMI[a]	2
11. logP[a]	1	30. Hbondacceptor[a]	3
12. $MR_{CM^{**}-3}$[a]	3	31. RotBonds[a]	3
13. MolRef[a]	3	32. Shadow-xlength[c]	1
14. Hf[a]	3	33. Shadow-ylength[c]	3
15. Fh20[a]	3	34.Shaodw-zlength[c]	1
16. Foct[a]	3	35. Shadow-YZ[c]	2
17. Apol[a]	3	36. Shadow-XZ[c]	2
18. Dipol-mag[a]	3	37.Shadow-nu[a]	1
19. Dipol-Y[a]	1		

[a] See the descriptions in Table 1. [b] E-state index. [c] Shadow indices.

thermodynamic, electronic, quantum mechanical descriptors, and E-state indices. Structural descriptors include the molecular weight (MW), the number of rotatable bonds (Rotbonds), and the number of hydrogen bonds (Hbond acceptor). The spatial descriptors[30] used involve radius of gyration (RadOfGyration), density, molecular surface area, principal moment of inertia (PMI), molecular volume, and shadow indices. The thermodynamic descriptors were taken describing the hydrophobic character (logP), refractivity (MolRef), heat of formation (Hf), and the dissolution free energy for water and octanol (Fh2o; Foct). The electronic descriptors[31] taken were concerning superdelocalizability (Sr), atomic polarizabilities (Apol), and the dipole moment (Dipole). Electrotopological-state indices (E-state indices) used involve S-aasC, S-aaN, S-aaCH, etc. Table 2 summarizes all molecular descriptors used as the original variables.

All these molecular descriptors were generated using the Cerius[2] 3.5 software system on a Silicon Graphics R3000 workstation.

The modified PSO, the classical stepwise regression, and PLS algorithm were written in Matlab 5.3 and run on a personal computer (Intel Pentium processor 4/1.5G Hz 256 MB RAM).

4. RESULTS AND DISCUSSION

4.1. Optimized Block-wise Variable Combination (OBVC) by the Modified PSO for Data Set 1. Molecular descriptors for Data Set 1 studied as original variables are listed, and the variables' allocations are also shown in Table 1. The minimum residual sum of the squares of the model varied with the value of *k*. The curve of the minimum residual sum of the squares of the model vs *k* is presented in Figure 2. The obvious plane section of the curve suggests *k* took the value of 2. Thus the original variables are allocated into two blocks. Based on PLS with the *F* statistic, we obtained five latent variables from the first block and one latent variable from the second block. These latent variables, which explain most of the variance of the block variables, were used to substitute the original individual 77 variables ultimately. Here one can see the way PLS with the introduced

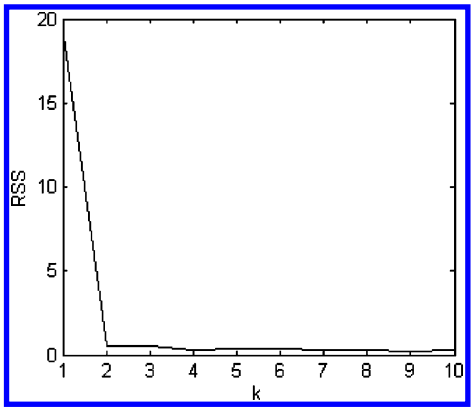

Figure 2. The curve of the minimum residual sum of the squares of the model for aromatic amine data vs *k*.

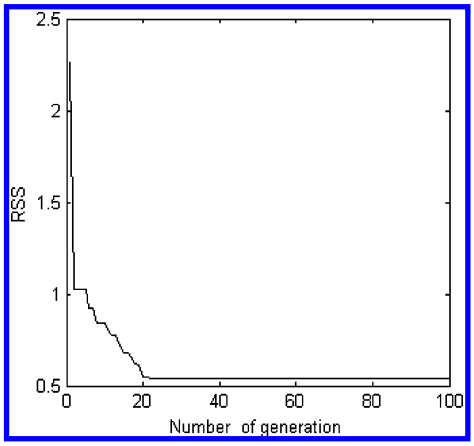

Figure 3. Convergence curve for aromatic amine data.

Table 3. Results of Data Set 1 Using OBVC by PSO Based on PLS Modeling

K[b]	T[c]		RSS[a]	RSS_{pred}[a]	R[a]	R_{pred}[a]
2	5	1	0.5399	1.9439	0.9923	0.9244

[a] RSS, RSS_{pred}: residual sum of the square for training and predicting set; R,R_{pred}: correlation coefficient for training and predicting set. [b] K: the number of blocks. [c] T: the numbers of latent variables extracted in each block.

F statistic works. The convergence curve of the residual sum of the squares of the model is presented in Figure 3. The results are summarized in Table 3. The correlation between the calculated and experimental values of BRR is shown in Figure 4. The results obtained by using OBVC by PSO followed by variable extraction based on PLS modeling are compared favorably with those acquired by stepwise regression. The results obtained by MLR after variable selection by stepwise regression gave an RSS of 0.8030 and 2.0425 for training and validation sets, respectively, and its correlation coefficient for the training set was 0.9886 and for the validation set was 0.9164.

4.2. Optimized Block-wise Variable Combination (OBVC) by the Modified PSO for Data Set 2. To further check the validity of the proposed methods, we applied the OBVC by the modified PSO for Data Set 2. The allocations of molecular descriptors for 4H-1,2,4-triazoles studied as original variables are presented in Table 2. According to the curve of the minimum residual sum of the squares of the model vs *k*, Figure 5, it seems that it is a comparatively good model when *k* takes the value of 3. Then the original

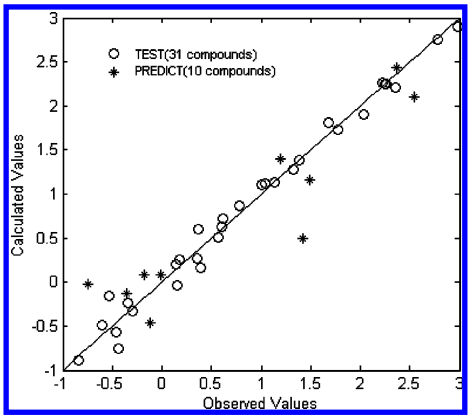

Figure 4. Calculated versus observed BRR using OBVC by PSO based on PLS modeling for aromatic amine data.

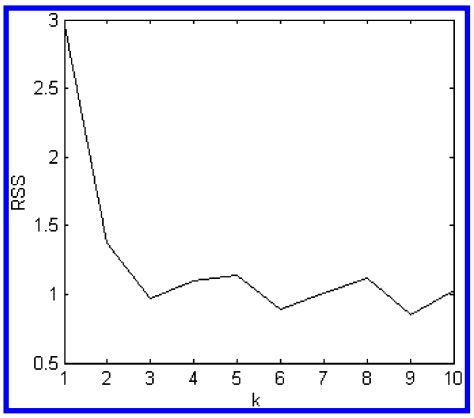

Figure 5. The curve of the minimum residual sum of the squares of the model for 4H-1,2,4-triazole data vs *k*.

Table 4. Results of Data Set 2 Using OBVC by PSO Based on PLS Modeling[a]

K	*T*			RSS	RSS_{pred}	*R*	R_{pred}
3	2	2	2	1.1892	1.6563	0.9725	0.9260

[a] See footnotes for Table 3.

variables are allocated into three blocks. Based on PLS modeling, two latent variables were extracted from the three blocks each automatically by the *F* statistic. The original individual 37 variables were eventually substituted by six latent variables. The results are summarized in Table 4. The convergence curve of the residual sum of the squares of the model is revealed in Figure 6. The correlation between the calculated and experimental values of $\log 1/IC_{50}$ is demonstrated in Figure 7. The results shown have been substantially improved as compared to those of stepwise regression. The results obtained by MLR after variable selection by stepwise regression gave an RSS of 2.1763 and 2.8951 for training and validation sets, respectively, and its correlation coefficient for the training set was 0.9490 and for the validation set was 0.8830.

From Tables 1 and 2, one can see that the similar descriptors having been considered as one group are allocated into different blocks. One may come to a conclusion that it is unadvisable to allocate the similar descriptors into one group subjectively. The explanatory variables finally obtained by the OBVC method are linear combinations of the original descriptors. The OBVC approach iteratively optimizes the descriptor blocks using the PSO algorithm. Consequently,

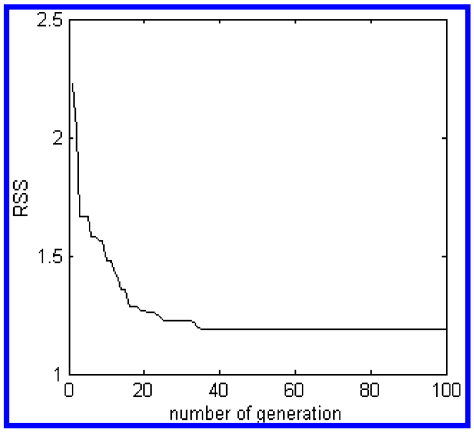

Figure 6. Convergence curve for 4H-1,2,4-triazole data.

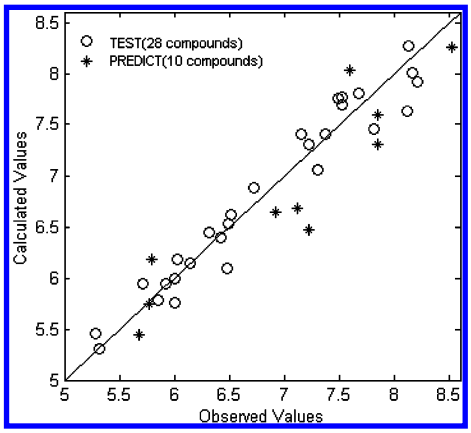

Figure 7. Calculated versus observed $\log 1/IC_{50}$ using OBVC by PSO based on PLS modeling for 4H-1,2,4-triazole data.

the linear combinations finally obtained, i.e., the explanatory variables, are the best ones for the PLS model over all possible combinations of descriptor blocks. This offers the possibility to refine the performance of the QSAR model in comparison to existing method. As was demonstrated in the present study, the OBVC approach offers much better performance than the stepwise variable selection method and the conventional PLS method.

The design of improved compounds using the OBVC approach is exactly consistent with all the existing QSAR methodologies. That is, to obtain improved compounds, certain modifications of the template compounds are first designed, then the descriptors of these modified compounds are computed, and the corresponding activity is predicted using the QSAR model to seek for the candidates of improved compounds. The only difference is that the QSAR model in the OBVC method is not built immediately using the descriptor but using the explanatory variables constructed as the linear combinations of the descriptors within a block preoptimized using known training compounds. With the explanatory variables thus obtained, the activity of these modified compounds can be predicted using the PLS model related to the explanatory variables to the dependent one.

With the *F* statistic introduced, there are few user-interfering operations in this process. The convergence processes for OBVC by the modified PSO can be examined in Figures 3 and 6. From these two figures, one can see that OBVC by the modified PSO can converge to a satisfactory solution in about 20 cycles for Data Set 1 and about 40 cycles for Data Set 2. The calculation results confirm that OBVC

by the modified PSO converges to the best solution quickly. The time required to perform the algorithm is only several minutes.

To inspect whether the OBVC method produces a correlation within the resulting meaningless data or not, we stochastically scrambled the y-block (biology) 10 times in Aromatic amine data to produce meaningless data sets. The correlation coefficients of the training sets for these y-scrambled Aromatic amine data are 0.5054, 0.5557, 0.5665, 0.5495, 0.5260, 0.8583, 0.8019, 0.5921, 0.7130, and 0.8387, respectively. Compared to the correlation coefficient 0.9923 for the original data, it is clear that the model obtained with the introduced OBVC algorithm is statistically stable and will not generate models due to chance correlation. The correlation coefficients of the training sets for Angiotensin II antagonists data with y-block stochastically scrambled 10 times were also calculated as follows: 0.6481, 0.8135, 0.6351, 0.5238, 0.4572, 0.4955, 0.5908, 0.5671, 0.8172, 0.6127. These correlation coefficients are also significantly smaller than that for the original data, 0.9725, indicating that the chance correlation associated with many explanatory variables was effectively circumvented using the OBVC approach. As a matter of fact, the OBVC method incorporates two strategies to eliminate chance correlation. One is the implementation of PLS algorithm in the construction of block-wise variable combination and the modeling of the relationship between the dependent variables and the resulting variable combinations. It is known that primary PLS latent variables seek to not only be most correlated to the dependent variable (y-block) but also account for most variance in the explanatory variables (x-block). Because chance correlation is always associated with small variations in x-block, the OBVC approach is expected to offer adequate resistance to chance correlation. The other strategy to avoid chance correlation is the introduction of an *F*-statistics to determine the number of PLS latent variables. This statistics requires that the latent variables to be included in the model should improve significantly the model error, which eliminates the risk of introducing less informative latent variables and mitigates the possibility of chance correlation. Therefore, it is expected that the OBVC method, though involving many descriptors, is not likely to generate QSAR models due to chance correlation.

Like most of the existing approaches, the OBVC method gives results dependent upon the descriptor choice. This is also an indication that the OBVC method is robust to chance correlation. Actually, it was found that the OBVC method generated a good model only with carefully selected original descriptors. However, even with good original descriptors, it is not always possible to obtain a good model with any fitting procedure. As was demonstrated in the manuscript, the stepwise regression procedure failed to produce a desirable model with the same choice of original descriptors. Without block-wise variable combination, the PLS method was also incapable of giving a model with satisfactory performance.

5. CONCLUSION

In the present study, the optimized block-wise variable combination is obtained by the modified particle swarm optimization based on the PLS model with the *F* statistic. Simultaneously an optimum PLS model is formulated based on a few latent variables extracted from the original variables. It has been demonstrated that the modified PSO is a useful tool for searching optimized variable combination which converges quickly toward the optimal position. Comparing with stepwise regression, the proposed methodology shows satisfactory prediction performance in the QSAR analysis of carcinogenicity of aromatic amines and antagonism of angiotensis II antagonists.

ACKNOWLEDGMENT

The work was financially supported by the National natural Science Foundation of China (Grant No. 20375012, 20105007, 20205005, 20435010).

REFERENCES AND NOTES

(1) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 77−84.

(2) 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.

(3) Agostinelli, C. Robust stepwise regression. *J. Appl. Stat.* **2002**, *29*(6), 825−840.

(4) 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.

(5) Kubinyi, H. Evolutionary variable selection in regression and PLS analyses. *J. Chemometrics* **1996**, *10*, 119−133.

(6) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.

(7) 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.

(8) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1218−1227.

(9) Du, Y. P.; Liang, Y. Z.; Li, B. Y.; Xu, C. J. Orthogonalization of Block Variables by Subspace-Projection for Quantitative Structure Property Relationship (QSPR) Research. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 993−1003.

(10) Kennedy, J.; Eberhart, R. Particle swarm optimization. In *IEEE Int. 1 Conf. On Neural Networks*; Perth: Australia, 1995; pp 1942−1948.

(11) Shi, Y.; Eberhart, R. A modified particle swarm optimizer. In *IEEE World Congress on Computational Intelligence*; 1998; p 69.

(12) Clerc, M.; Kennedy, J. The particle swarm−explosion, stability, and convergence in a multidimensional complex space. *IEEE transactions on evolutionary computation*; 2002; Vol. 6, pp 58.

(13) Shi, Y.; Eberhart, R. Fuzzy adaptive particle swarm optimization. In *Proc congress on evolutionary computation*; 2001.

(14) Shen, Q.; Jiang, J. H.; Yu, R. Q. Modified Particle Swarm Optimization Algorithm for Variable Selection in MLR and PLS modeling: QSAR Studies of Antagonism of Angiotensin II Antagonists. *Eur. Pharm. Sci.* Accepted for publication.

(15) Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons: New York.

(16) Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A. Quantitative Structure−Activity Relationships of Mutagenic and Carcinogenic Aromatic Amines. *Chem Rev.* **2000**, *100*, 3697−3714.

(17) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.

(18) Hall, L. H.; Kier, L. B. The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−78.

(19) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: An Atom Index for QSAR. *Quant. Struct-Act. Relat.* **1991**, *10*, 43−48.

(20) Kier, L. B.; Hall, L. H.; Frazer, J. W. An Index of Electrotopological State for Atoms in Molecules. *J. Math. Chem*. **1991**, *7*, 229−237.

(21) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(22) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure−activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(23) Dewar, M. J. S.; Thiel, W. Ground states of molecules. 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc.* **1977**, *99*, 4899−4907.
(24) Hopfinger, A. J.; et al. *Safe handling of chemical carcinogens, mutagens, teratogens and highly toxic substances*; Ann Arbor Press: Ann Arbor, MI, 1980; p 385.
(25) Gasteiger, J.; Marsali, M. Iterative partial equalization of orbital electronegativity−−a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219.
(26) Stanton, D. T.; Jurs, P. C. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure−property relationship studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(27) Bonchev, D. *Information theoretic indices for characterization of chemical structures*; Research Studies Press: Chichester, U.K., 1983; p 249.
(28) Kurup, A.; Garg, R.; Carini, D. J.; Hansch, C. Comparative QSAR: angiotensin II antagonists. *Chem. Rev.* **2001**, 2727−2750.
(29) Ashton, W. T.; Cantone, C. L.; Chang, L. L. Nonpeptide angiotensin II antagonists derived from 4H-1,2,4-triazoles and 3H-tmidazo triazoles. *J. Med. Chem.* **1993**, *36*, 591−609.
(30) Rohrbaugh, R. H.; Jurs, P. C. Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. *Anal. Chim. Acta* **1987**, *199*, 99−109.
(31) *Cerius² QSAR+*; Molecular Simulations Inc.: San Diego, 1997.

CI049890I