Optimized Partition of Minimum Spanning Tree for Piecewise Modeling by Particle Swarm Algorithm. QSAR Studies of Antagonism of Angiotensin II Antagonists

Qi Shen, Jian-Hui Jiang, Chen-Xu Jiao, Shuang-Yan Huan, 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 December 17, 2003

In quantitative structure−activity relationship (QSAR) modeling, when compounds in a training set exhibit a significant structural distinction between each other, in particular when chemicals of biological interest interacting on the receptor involve a different mechanism, it might be difficult to construct a single linear model for the whole population of compounds of interest with desired residuals. Developing a piecewise linear local model can be effective to circumvent the aforementioned problem. In this paper, piecewise modeling by the particle swarm optimization (PMPSO) approach is applied to QSAR study. The minimum spanning tree is used for clustering all compounds in the training set to form a tree, and the modified discrete PSO is applied to divide the tree to find satisfactory piecewise linear models. A new objective function is formulated for searching the appropriate piecewise linear models. The proposed PMPSO algorithm was used to predict the antagonism of angiotensin II. The results demonstrated that PMPSO is useful for improvement of the performance of regression models.

1. INTRODUCTION

Quantitative structure−activity relationship (QSAR) studies focus on construction of QSAR models that relate the molecular structures of chemical compounds to a certain biological activity of interest. In cases of a training set consisting of structurally similar analogues, QSAR models can often be improved by variable selection. If compounds with multiple templates in a training set exhibit a significant structural distinction between each other, in particular the chemicals of biological interest interact on the object receptor with a different mechanism, performing variable selection alone might not be able to find the desired unique model. Here, the properties of compounds in a training set are of fundamental importance in regard to the quality and performance of regression models. In such a case it might be difficult to construct a single linear model for the whole population of compounds of interest with desired residuals.[1−3] To circumvent the aforementioned problems, besides the nonlinear approaches using sophisticated nonlinear functions, an alternative approach is to find multiple models by splitting the whole data set into subsets with desired linearity in each group.[4,5] Representation of a nonlinear system using multiple models can be achieved by approximating the nonlinearity with piecewise local linear models.[6−8] Most piecewise linear approaches in the literature are based on the natural ordering of the observations, for example, with the help of a time parameter *t*, and desired linearity was obtained by splitting the data set into subsets corresponding to intervals for parameter *t*. In QSAR research unfortunately no such ordering is available. In addition, the process of constructing piecewise linear models and assigning predicted compounds in the test set to appropriate models obtained by data set splitting method heavily relied on the expertise of the researcher and can hardly be fully automated.

Piecewise models for the compounds involved can be searched using clustering methodology. A training set containing structurally diverse compounds can be grouped into clusters first and then each cluster used to fit a model. In contrast to the conventional cluster analysis, with piecewise modeling attention needs to be paid to not only the closeness of samples in space but also the linearity within each group. As it may lead to overfitting when the number of compounds is too small compared to the number of variables, the stability of each submodel needs to be taken care of.

The goal of splitting methodologies is to optimize the subsets that each can be modeled well by a linear model. Splitting the data set to subsets is an optimization problem. Particle swarm optimization (PSO),[9−12] a relatively new optimization technique originated from simulation of a simplified social system, is useful for solving the piecewise linear model optimization problem. Similar to the genetic algorithm (GA) and evolution algorithm (EA), PSO is a population-based optimization tool which searches for optima by updating generations. However, unlike GA and EA, PSO possesses no evolution operators such as crossover and mutation. Compared to GA and EA, PSO has the advantage of being conceptually very simple, requiring low computation costs and few parameters to adjust. A modified discrete PSO algorithm[13] has been proposed in our previous study to select variables in MLR and PLS modeling with satisfactory performance. In the present study, we proposed a procedure based on a minimum spanning tree and PSO algorithms that can be used to split a training data set into subsets with desired linearity in each class automatically without interference of the user. Determining which model should be applied to a given test set compound is based on the Euclidean

* Corresponding author phone: +86-731-8821577; fax 86 731 8822782; e-mail: rqyu@hnu.net.cn.

10.1021/ci034292+ CCC: $27.50
Published on Web 08/25/2004

distance. It is consistent with the principle of building multiple models. A new objective function is formulated to determine the appropriate multiple models. The formulation and corresponding programming flowchart are presented in details in the paper. As an example of application of the proposed multiple model optimization algorithm, antagonism of angiotensin II was predicted using piecewise linear models. The results have been demonstrated that the proposed method is useful for improving the performance of regression models.

2. THEORY

2.1. Minimum Spanning Tree. Piecewise modeling by particle swarm optimization was based on cluster analysis. Minimum spanning tree (MST)[14,15] is one of the well-known techniques of hierarchical clustering. First, the minimum spanning tree algorithm was used to connect N objects or points. MST uses the graph theory concept to connect a set of points and minimize the total length of connecting lines. A spanning tree of a graph is a set of $N - 1$ edges that connect all the N objects of the graph without cycles. The MST is the set of edges with minimum sum of the lengths over the $N - 1$ edges. In our algorithm, the length associated with each edge is the Euclidean distance between the two connected objects.

Numerous algorithms have been developed to compute a MST. The Kruskal approach,[15] one of the algorithms usually used to find the MST connecting N objects, was used in this study. The Kruskal algorithm starts by calculating the distance between each pair of points and arranging the pairs by the distance between them from closest to farthest. One connects the first pair of points with a line involving the two nearest objects. Then the following pairs on the list are considered. In each step the two nearest objects are connected, provided they are not both in the same tree already formed. If both points are in the same tree already formed, ignore the pair and move on. When both objects are not in any tree previously formed, they are connected and constitute a new tree. Thus, a number of separated trees can be formed. When the two connected points are in different trees, the two trees merge. The algorithm continues until all points are connected in a unique tree. MST algorithm chooses at each step the shortest edge to add to the tree.

2.2. Modified Particle Swarm Optimization. PSO,[9−12] developed by Eberhart and Kennedy in 1995, is a stochastic global optimization technique inspired by social behavior of bird flocking. The algorithm models exploration of a problem space by a population of individuals or particles. In PSO each single solution is a particle in the search space. Each individual in PSO flies in the search space with a velocity that is dynamically adjusted according to its own flying experience as well as that of its companions. In 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. PSO is initialized with a group of random particles. Each particle is treated as a point in a D-dimensional space. The ith particle is represented as $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iD})$. The best previous position of the ith particle that gives the best-fit value is represented as $\mathbf{p}_i = (p_{i1}, p_{i2}, ..., p_{iD})$. The best particle among all particles in the population is represented by $\mathbf{p}_g = (p_{g1}, p_{g2}, ..., p_{gD})$. Velocity, the rate of position change for particle I, is represented as $\mathbf{v}_i = (v_{i1}, v_{i2}, ..., v_{iD})$. In every iteration each particle is updated by following the two best values.

For a discrete problem expressed in binary notation, a particle moves in a search space restricted to 0 or 1 on each dimension. In a binary problem, updating a particle represents changes of a bit that should be in either 1 or 0 state and the velocity represents the probability of bit x_{id} taking the value 1 or 0.

According to the information sharing mechanism of PSO, a modified discrete PSO[13] was proposed as follows. The velocity v_{id} of every individual is a random number in the range (0,1). The resulting change in position then is defined by the following rule

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

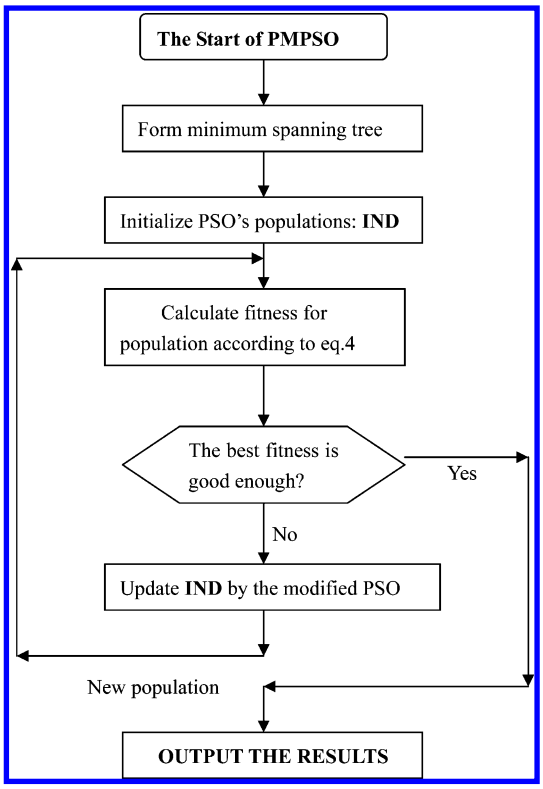
$$\text{If } (a < v_{id} \le (1 + a)/2), \text{ then } x_{id}(\text{new}) = p_{id} \quad (2)$$

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

where a is a random value in the range of (0,1) named static probability. In this study static probability $a = 0.5$ Though the velocity in the modified discrete PSO is different from that in the continuous version of PSO, the information sharing mechanism and updating model of a particle by following the two best positions is the same in two PSO versions. To circumvent convergence to local optima and improve the ability of the modified PSO algorithm to overleap local optima, 10% of particles are forced to fly randomly not following the two best particles. If the minimum error criterion is attained or the number of cycles reaches a user-defined limit, the algorithm is terminated.

2.3. Piecewise Modeling by Particle Swarm Optimization (PMPSO). If compounds in a training set exhibit a significant structural distinction from each other, it might be difficult to construct a single linear model for the whole population of compounds of interest with the desired residual. An efficient scheme is to split the single model into multiple models with improved linearity in each model. MST is used for clustering all compounds in a training set to form a tree, and the modified discrete PSO is applied to divide the tree to find satisfactory piecewise linear models. Dividing the tree is not in view of distance between compounds but based on fitting each model. Splitting MST should give improved linearity in each group, and the information in both independent and dependent variables are all used. Each point in MST stands for an object or compound in the training set. Each edge that connects two objects stands for a site that may be divided. In the modified discrete PSO, each particle is encoded to a string of binary bits associated with the number of edges, which makes up the MST together with all its edges. A bit of 1 in a particle represents segmentation of the corresponding edge. If there are N edges be divided, the training set will be split into $N + 1$ subsets, i.e., there will be $N + 1$ models in the training set. The linear models for each of the groups created by the splits are encoded in the PSO algorithm. Models are calculated for each of the groups created by the splits present in each individual in the PSO algorithm. In this algorithm the number of models

Scheme 1 Chart of the PMPSO Scheme



involved in computation is automatically adjusted. The piecewise modeling by particle swarm optimization (PMPSO) is described as follows.

Step 1: Scale the value of the data set into (0.0, 1.0). Form MST using all compounds in the training set.

Step 2: Randomly initialize all the initial strings **IND** in modified discrete PSO with an appropriate size of population. **IND** are strings of binary bits corresponding to each edge in **MST**.

Step 3: Calculate the fitness function of individuals corresponding to models in the training set. If the best object function of the generation fulfills the end condition, the training is stopped with the results output; otherwise, go to the next step.

Step 4: Update the **IND** population according to the modified discrete PSO.

Step 5: Go back to the third step to calculate the fitness of the renewed population

The Euclidean distances between the new compounds and compounds in the training set are used to decide the models that the compounds to be predicted are belong to. It is consistent with the distances in MST. If compound A in the test set is nearest to compound B in the training set, the model which includes B is used to predict A. The PMPSO scheme is presented in Scheme 1.

2.4. Fitness Function. In PMPSO, the performance of each particle is measured according to a predefined fitness function. Splitting the whole data set into subsets should necessarily improve the performance of each model. According to experience, if the number of descriptors (p) is too large compared to the number of observations (n), this may deteriorate the performance of QSAR modeling. In light of these requirements, we can formulate the objection function whose minimization will generate optimum multiple models. The fitness function of PMPSO is evaluated based on two aspects: the accuracy of multiple models and the dimension of each model. The fitness is defined as

$$\text{fitness} = \sum_{i=1}^{N}\{[\sum_{j=1}^{n}(Y_j - YP_j)^2][1 + b*e^{(-n/p)}/(1 + e^{(-n/p)})]\} \quad (4)$$

where N is the number of models, n is the number of observations or compounds in each model, p is the number of variables, and b is the weighting coefficient between the accuracy and dimension of each model to penalize cases with too few samples. According to experience b is set to 1. Y_j and YP_j are, respectively, the experimental and calculated values of the ith sample. The first term of the right side of eq 4 is the sum of squared residuals (RSS), which is defined as the accuracy of each model, and the dimension of each model is restricted by a sigmoidal function (the second term of the right side of eq 4). The sigmoid function in the second term on the right side of eq 4 is used for limiting the dimension of each model. If the number of descriptors (p) is too large or too small compared to the number of observations (n), the sigmoid function will be much larger than that with the appropriate number of descriptors.

3. ANGIOTENSIN II ANTAGONISTS DATA

A set of 85 1,2,4-triazoles[16] as angiotensin II antagonists, which were synthesized and evaluated for their antagonism against angiotensin II by Ashton et al., was used to test the performance of PMPSO in QSAR studies. The antagonism against angiotensin is expressed as IC_{50}, the molar concentration of the compound causing 50% antagonism of angiotensin II.

The data set of 85 1,2,4-triazoles was randomly divided into two groups with 74 compounds used as the training set for developing regression models and the remaining 11 compounds used as the validation set in the prediction of antagonism against angiotensin II.

A series of molecular descriptors was calculated for 1,2,4-triazoles including spatial, structural, electronic, and thermodynamic descriptors as well as *E*-state indices. In piecewise modeling each sample is described by the following five parameters: principal moment of inertia (PMI), hydrophobic character (AlogP: logarithm of the partition coefficient in octano/water),[17] molar refractivity (MolRef:), lowest unoccupied molecular orbital energy (LUMO), and electrotopological-state indices[18,19] (S-sOH). The *E*-state index S-sOH for the hydroxide radical represents the electron accessibility associated with it. In the symbol S-sOH, 'S' represents the electronic topological state of atom, 's' the single bond of the group, and 'OH' the hydroxyl radical.

The PMPSO algorithm was programmed in Matlab 6.0 and run on a personal computer (Intel Pentium processor 733 MHz, 128 MB RAM).

4. RESULTS AND DISCUSSION

4.1. QSAR Study by Modeling as a Whole Data Set. Antagonism of 85 1,2,4-triazoles data were taken to evaluate the PMPSO algorithm. For comparison with PMPSO, the data of antagonism were first analyzed as a whole data set by MLR modeling. The variable selection analysis was

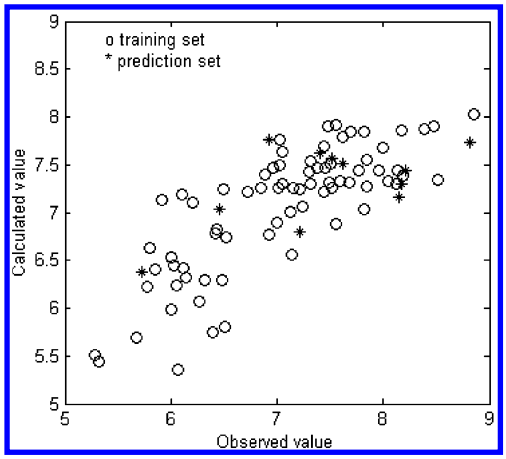


Figure 1. Calculated versus observed log(1/IC_{50}) by MLR modeling using a whole data set.

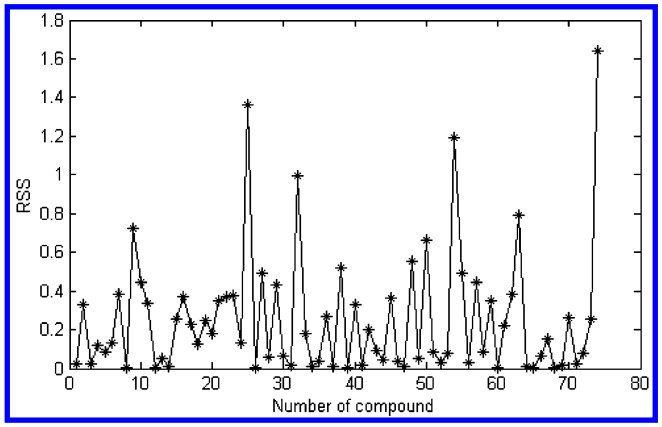


Figure 2. RSS by MLR modeling using a whole data set.

performed to improve the MLR model. The best model contains five variables during GAs search. The five variables are PMI, AlogP, MolRef, LUMO, and S-sOH. The correlation coefficients (*R*) for the training set and validation set were 0.7907 and 0.6133, respectively. The correlation between the calculated and experimental values of antagonism is shown in Figure 1. The sum of squared residuals (RSS) was 18.19. The squared residuals for the whole training data set are shown in Figure 2. As shown in Figures 1 and 2, the correlation was rather poor and modeling error was quite high. Using a single model cannot describe the data set successfully, even after careful variable selection.

4.2. QSAR Study of Piecewise Modeling by Particle Swarm Optimization. Optimization of the MLR model by variable selection analysis only is inadequate to obtain a good unique model. The reason is associated with the difficulty in constructing a single linear model for the whole data set with desired residuals. To improve the QSAR modeling of antagonism, PMPSO was used to split the data set into multiple models. The best five variables selected by GA search were used to test the performance of PMPSO. In PMPSO, MST is used for clustering all compounds in a training set to form a tree and the modified discrete PSO is applied to divide the tree to find satisfactory piecewise models. The population size of PSO was selected as 20, and the weighting coefficient *b* in eq 4 was set to 1.

Data set splitting by PMPSO should improve the linearity in each group, and the number of variables should not be too large compared to the number of observations. PMPSO

Table 1. Result of QSAR Analysis

method/ data set	*R* (correlation coefficient)		RSS (sum of squared residual)	
	method 1[a]	method 2[b]	method 1[a]	method 2[b]
subset 1	0.7637	0.6684	6.0997	8.7239
subset 2	0.6943	0.6543	4.9859	5.8143
subset 3	0.9249	0.7955	0.5840	4.2742
whole data set	0.8675	0.7958	11.6696	18.1905
test set	0.7158	0.6264	4.7692	5.2220

[a] Method 1: PMPSO. [b] Method 2: QSAR study by modeling as a whole data set.

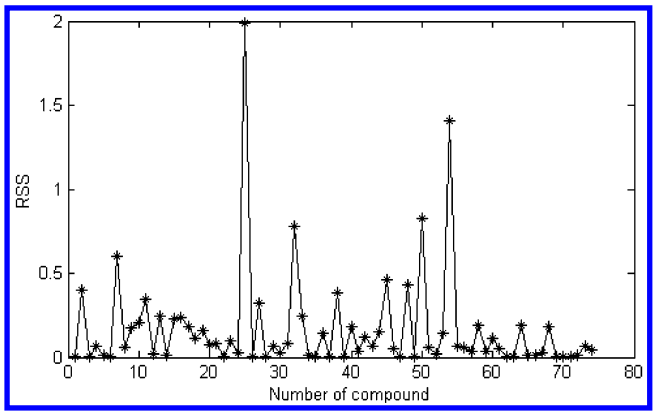


Figure 3. RSS by PMPSO when the whole data was split into three subgroups.

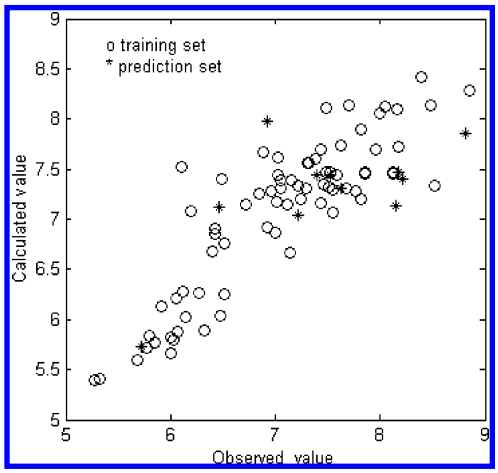


Figure 4. Calculated versus observed log(1/IC_{50}) using PMPSO.

divided the whole data set into three linear subsets. The three subsets contain 32, 26, and 16 compounds, respectively. The results of PMPSO are listed in Table 1. RSS of compounds in each subset modeled by PMPSO were smaller compared to RSS of these compounds in a single model. *R* in each submodel by PMPSO was higher than that in a model as a whole data set. RSS was reduced from 18.19 to 11.67 for all compounds by PMPSO. The squared residuals for all observations by PMPSO are shown in Figure 3. The corresponding residuals by PMPSO were smaller than those in a single model by MLR. The correlation between the calculated and experimental values of antagonism is shown in Figure 4. Using the same five variables, the correlation coefficients for the training set and validation set were improved to 0.8675 and 0.7158, respectively. Prediction of Compounds in the test set were predicted using the submodel selected by the Euclidean distance. *R* for the test set increased from 0.6264 by a single model to 0.7158 by PMPSO. A comparison of Figures 1 and 2 and Figures 3 and 4 shows

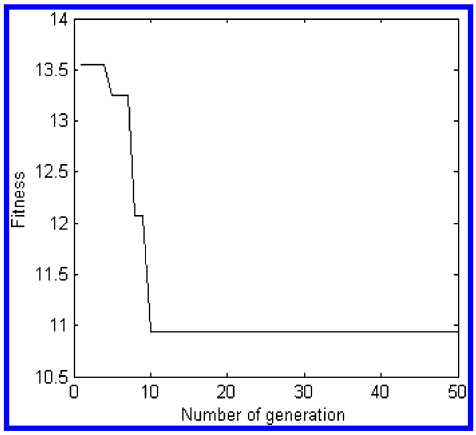


Figure 5. Convergence curves for PMPSO.

that better results were obtained from multiple models by the PMPSO algorithm than by MLR as a whole data set.

If compounds in a subset were predicted by an incorrect subset model, *R* was significantly lower and RSS was higher than that by individual models. The RSS for the test set is 4.7692 by PMPSO. If using those incorrect subset models predicted these test compounds, the RSS for the test set will be higher than 6. This indicates that the piecewise modeling is really useful for improving the QSAR model. Performing PMPSO has been found to be useful in developing models.

The convergence processes for PMPSO are examined in Figure 5. From Figure 5 one can see that PMPSO can converge to a satisfactory solution in about 11 cycles. Experimental results confirm that the PMPSO converges to the best solution quickly. The time required to perform the PMPSO is only several minutes.

Parameter *b* in the fitness function (eq 4) is the weighting coefficient between model accuracy and dimension of each model to penalize few samples. The smaller the value of the parameter *b* is, the fewer compounds in subgroups. This may lead to overfitting as the number of compounds is too small compared to the number of variables. On the other hand, a large value of parameter *b* is favorable for avoiding overfitting but may not improve the QSAR model effectively. Accordingly, *b* was set to 1 by experience to keep balance between the accuracy and dimension of the model.

5. CONCLUSION

Developing piecewise modeling is useful for improving QSAR modeling. In this paper, a minimum spanning tree was used for clustering all compounds in a training set to form a tree and the modified discrete PSO was applied to divide the tree to find satisfied piecewise linear models. A new objective function was formulated to determine the appropriate piecewise models. Antagonisms of angiotensin II were predicted by the proposed piecewise modeling. The results demonstrate that the proposed method is useful for improvement of the performance of regression models.

ACKNOWLEDGMENT

The work was financially supported by the National Natural Science Foundation of China (Grant Nos. 20105007, 20205005, 20375012)

REFERENCES AND NOTES

(1) Cho, S. J.; Hermsmeier, M. A. Genetic Algorithm Guided Selection: Variable Selection and Subset Selection. *J. Chem. Inf. Comput.* **2002**, *42*, 927−936.

(2) Nasser, A. M. B.; Jiang, J. H.; Liang, Y. Z.; Yu, R. Q. Piece-wise Quasi-linear Modeling in QSAR and Analytical Calibration Based on Linear Substructures Detected by Genetic Algorithm. *Chemom. Intell. Lab. Syst.* **2003**, in press.

(3) Du, Y. P.; Liang, Y. Z.; Yun, D. Data Mining for Seeking an Accurate Quantitative Relationship between Molecular Structure and GC Retention Indices of Alkenes by Projection Pursuit. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1283−1292.

(4) Tormod, N.; Tomas, I. Splitting of Calibration Data by Cluster Analysis. *J. Chemom.* **1991**, *5*, 49−65.

(5) Tormod, N. Multivariate Calibration When Data are Split into Subsets *J. Chemom.* **1991**, *5*, 487−501.

(6) Leyla, O.; Mayuresh, V. K.; Christos, G. Model Predictive Control of Nonlinear Systems Using Piecewise Linear Models. *Comput. Chem. Eng.* **2002**, *24*, 793−799.

(7) Lin, J. N.; Rolf, U. Explicit Piecewise-Linear Models. *IEEE Trans. Circuit Syst.* **1994**, *12*, 931.

(8) Morimoto, Y.; Madarame, H. Piecewise Linear Model for water column oscillator simulating reactor safety system. *Int. J. Nonlinear Mech.* **2003**, *38*, 213−223.

(9) Kennedy, J.; Eberhart, R. Particle swarm optimization. In *IEEE International 1 Conference On Neural Networks*, Perth, Australia, 1995; pp 1942−1948.

(10) Shi, Y.; Eberhart, R. A modified particle swarm optimizer. *IEEE World Congress Comput. Intell.* **1998**, 69−73.

(11) Clerc, M.; Kennedy, J. The particle swarm-explosion, stability and convergence in a multidimensional complex space. *IEEE Trans. Evol. Comput.* **2002**, *6*, 58−64.

(12) Shi, Y.; Eberhart, R. Fuzzy adaptive particle swarm optimization. *Proc. Congress Evol. Comput.* 2001.

(13) Shen, Q.; Jing, 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.*, submitted for publication

(14) Prim, R. C. Shortest connection matrix network and some generalizations. *Bell Syst. Techn. J.* **1957**, 36.

(15) Kruskal, J. B. On the shortest spanning tree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.* **1956**, *7*, 48−50.

(16) 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.

(17) 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.

(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.

(19) 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.

CI034292+