ARTICLES

Piecewise Hypersphere Modeling by Particle Swarm Optimization in QSAR Studies of Bioactivities of Chemical Compounds

Wei-Qi Lin, Jian-Hui Jiang, Qi Shen, Hai-Long Wu, 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 November 29, 2004

As the structural diversity in a quantitative structure−activity relationship (QSAR) model increases, constructing a good model becomes increasingly difficult, and simply performing variable selection might not be sufficient to improve the model quality to make it practically usable. To combat this difficulty, an approach based on piecewise hypersphere modeling by particle swarm optimization (PHMPSO) is developed in this paper. It treats the linear models describing the sought-for subsets as hyperspheres which have different radii in the data space. According to the attribute of each hypersphere, all compounds in the training set are allocated to hyperspheres to construct submodels, and particle swarm optimization (PSO) is applied to search the optimal hyperspheres for finding satisfactory piecewise linear models. A new objective function is formulated to determine the appropriate piecewise models. The performance is assessed using three QSAR data sets. Experimental results have shown the good performance of this technique in improving the QSAR modeling.

1. INTRODUCTION

Quantitative structure−activity relationship (QSAR) studies relating chemical structure to biological activity play an important role in drug design. In traditional medicinal chemistry, the optimization of a QSAR model is performed at the descriptor level,[1−8] and the compounds typically found in such a model tend to consist of structurally similar analogues. As the structural diversity in the training set increases, the variable selection analysis alone may be inadequate to develop a good QSAR model. In this case, the quality of the model may depend less on the types of variables present and more on the types of compounds present in the data set. Very recently Cho and co-workers[9] reported a genetic algorithm guided selection to optimize the encoded variables that include both descriptors and compound subsets.

In the present study, a new approach based on piecewise hypersphere modeling by particle swarm optimization (PHMPSO) is introduced to split the data set into subsets which are described by linear models with a desired error level. Particle swarm optimization (PSO) developed by Eberhart and Kennedy[10−14] in 1995 is a stochastic global optimization technique inspired by the social behavior of bird flocking. PSO has been found to be robust and fast in solving a lot of optimization problems. The recent years have seen a growing interest in the application of PSO as an optimization technique for chemical problems. A modified discrete version of PSO proposed in this laboratory was used to optimize partitions of minimum spanning trees for piecewise modeling.[15] A hybridized particle swarm optimization approach was also developed for neural network structure training.[16] In this study the training set consisting of compounds that exhibit significant structural distinction between each other are grouped into clusters first based on probability distances. Each cluster used to construct a submodel will be considered as a hypersphere in the augmented data space from the geometrical point of view. The PSO algorithm is invoked to determine the optimal hyperspheres used to construct a submodel. Determining which model should be applied to a given test set compound is based on probability distance to hypersphere cores rather than the Euclidean distance. A new objective function is formulated to determine the appropriate multiple models. The proposed multiple model optimization algorithm has been first used for the Hansch data set to examine its ability and then to predict inhibitory activities of inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase, and it was applied to the data set of nonpeptide angiotensin II antagonists. The results have shown that the proposed method could be useful in improving a QSAR model.

2. THEORY

2.1. Piecewise Hypersphere Modeling. Piecewise hypersphere modeling for the compounds involved can be achieved by cluster analysis. Compounds in the training sets are grouped into hard clusters based on a probability distance measure to hypersphere cores that is a weighted Euclidean distance. The term 'hard' means that each sample belongs to only one cluster. Each cluster will be considered as a hypersphere in the augmented data space. As the volumes of clusters might be different, the hyperspheres will have different radii and each one has its own characteristic attribute. According to the attribute of each hypersphere, all

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

10.1021/ci049642m CCC: $30.25 © 2005 American Chemical Society
Published on Web 03/11/2005

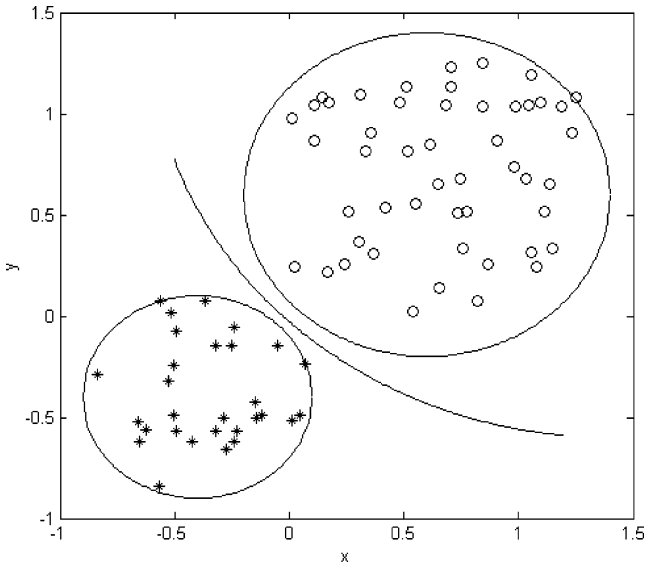

Figure 1. The sketch map of the formation of hyperspheres.

compounds in the training set are allocated to hyperspheres. The hyperspheres are formed in a high-dimensional space which generally could not be visualized. Figure 1 is the sketch map which presents the formation of hyperspheres graphically in a two-dimensional space. First, calculate the probability distance between the data point and each hypersphere core, which is the Euclidean distance calculated using selected descriptors between the data point and hypersphere core multiplied by the corresponding hypersphere weight. The hypersphere weight, defined as the reciprocal of the square of the hypersphere radius, takes into consideration the influence of the volume variation of hyperspheres on the allocation of a particular sample point to different clusters. If the probability distance between data point *j* and hypersphere *k* is a minimum, the data point *j* is allocated to hypersphere *k*. In this way, all data points are allocated to K hyperspheres. Such a series of hyperspheres is a single solution in the optimization.

2.2. Particle Swarm Optimization. Particle swarm optimization (PSO) introduced as an optimization technique by Eberhart and Kennedy in 1995 simulates social behavior among individuals (particles) "flying" through a multidimensional search space, and each particle keeps track of its space coordinates which are associated with the best solution (fitness) it has achieved so far. 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. 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 location combining 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. After finding the aforementioned two best values, the particle updates its velocity and positions according to the following equations

$$v_{id}(\text{new}) = w*v_{id}(\text{old}) + c_1 * r_1 * (p_{id} - x_{id}) + c_2 * r_2 * (p_{gd} - x_{id}) \quad (1)$$

$$x_{id}(\text{new}) = x_{id}(\text{old}) + \mu*v_{id}(\text{new}) \quad (2)$$

where w is an inertia weight which is brought into eq 1 to balance the global search and local search, c_1 and c_2 are two positive constants named as learning factors, and from experience both constants take the integer value 2, and r_1 and r_2 are random numbers in the range of (0,1). The time parameter μ in eq 2 determines the different flying time for each particle. Equation 1 is used to calculate the particle's new velocity according to its previous velocity and the distances of its current position from its own best position and the group's best position. Then the particle flies toward a new position according to eq 2. Such an adjustment of the particle's movement through the space causes it to search around the two best positions. If the minimum error criterion is attained or the number of cycles reaches a user-defined limit, the algorithm is terminated. To circumvent convergence to local optima and improve the ability of the PSO algorithm to overleap local optima, 10 percent of the particles are forced to fly randomly not following the two best positions.

2.3. Piecewise Hypersphere Modeling by Particle Swarm Optimization. As the structural diversity in a QSAR training set increases, constructing a good model becomes increasingly difficult, and simply performing variable selection might not improve the quality of the model. To combat this difficulty, an efficient scheme is to find multiple models by splitting the whole data set into subsets with improved linearity in each model. In this algorithm, compounds in the training set are grouped into hard clusters based on a probability distance measure to hypersphere cores. Each data point in a hypersphere stands for an object or compound in the training set. In PSO, each single solution is a particle in the search space, and each particle is encoded to a real string, the bits of which stand for the cores and radii of a series of hyperspheres, respectively. The piecewise hypersphere modeling by particle swarm optimization (PHMPSO) is described as follows.

Step 1. Normalize the data set. And randomly initialize all the hypersphere cores and radii in the PSO algorithm with an appropriate size of population. Calculate probability distances between data points and hypersphere cores. Then the data point is allocated to the hypersphere with the minimum probability distance to the sphere's core.

Step 2. Evaluate the fitness function of individuals in the current population. 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 PSO algorithm. Based on the renewed population, reallocate the data points to hyperspheres.

Step 4. Go back to the second step.

Thus the optimal hyperspheres used to construct submodels are determined ultimately. The prediction of compounds in the test set uses the submodel selected by probability distances to these hypersphere cores.

2.4. Fitness Function. In this algorithm, the performance of each particle is evaluated by a predefined fitness function. Splitting the whole data set into subsets should necessarily reduce the residuals from linear fitting within each subset and at the same time pay attention to the compactness of clusters in the data space. When the number of compounds is too few compared to the number of variables, it may lead to overfitting, so the stability of each submodel needs to be taken care of. In view of these requirements, the fitness function is defined as follows

$$\text{fitness} = [\sum_{j=1}^{N}(Y_j - YP_j)^2]*\left(1+\rho*\left[\frac{\sum_{k=1}^{K}\sum_{x_j\in\Omega_k}(||x_j - C_k||^2 * w_k)}{\sum_{j=1}^{N}||x_j - m||^2}\right]\right) \quad (3)$$

where N is the number of samples in the model, K is the number of the hyperspheres, i.e., the number of submodels, and Ω_k represents the whole set of compounds in hypersphere *k*. C_k is the core of hypersphere *k*, w_k is the weight of the corresponding hypersphere *k*, m is the mean of all the data points, and ρ is the weighting coefficient to keep a balance between accuracy and the compactness of clusters in the data space. By experience ρ takes the value of 0.1. Y_j and YP_j are, respectively, the experimental and calculated bioactivity values of the jth sample. The first term of the right side of eq 3 is the sum of squared residuals (RSS), which defines the accuracy of each model. The compactness of clusters is controlled by the second term of the right side of eq 3.

3. DATA SETS

3.1. Hansch Data Set. The Hansch data set[17,18] contains 111 2,4-diamino-5-(3,4-dichlorophenyl)-6-substituted pyrimidines acting as inhibitors of dihydrofolate reductase. Both the activities and descriptors are used as originally reported in ref 17 in order to compare with the published results; variable numbers 1−13 refer to variables I-1, I-2, I-4, I-7, I-8, I-9, I-10, I-13, I-15, I-17, I-20, I-4·I-8 and I-8·I-17 of the original publication (*k*=13; Table 1 of ref 17).

3.2. Inhibitors of the Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinase Data. In the present study sixty-one inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase 4-(X-phenylamino-)-Y-quinazoline[19] were used to test the performance of PHMPSO in QSAR studies. The molecular basic structure of inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase is presented in Figure 2. The detailed structural formulas of the compounds are listed in Table 1of the Supporting Information. The activity is expressed in molar concentration. I_{50} is the 50% inhibition concentration. The data set was stochastically divided into two groups. Forty-six compounds were used as the training set for developing regression models, while the remaining fifteen compounds were used as the prediction set. Each sample is described by the top 5 best descriptors reported in the reference. The five descriptors are as follows: the hydrophobic character (ClogP, logarithm of octanol/water partition coefficient),[20] indicator variable I (I=1 is an indication of the presence of 6,7-di-OMe derivatives), steric parameter $B1_{Y,7}$ and $B1_{X,3}$ (positive $B1_{Y,7}$ shows that Y-substituents at 7-position enhance the activity and it holds true to $B1_{X,3}$ for X-substituentes at the 3-position of the 4-phenylamino moiety),[19] the electronic descriptor σ^-_Y (negative σ^-_Y indicates that electron-donating groups such as Y-substituents enhance the activity).[19]

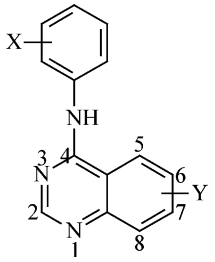

Figure 2. The parent structure of 4-(X-phenylamino)-Y-quinazoline.

Table 1. Results of QSAR Analysis of the Hansch Data Set by PHMPSO When the Whole Data Were Split into Two Subgroups Compared with that Obtained by MLR Modeling Using a Whole Data Set

	R (correlation coefficient)		RSS (sum of squared residual)	
data set	method 1[a]	method 2[b]	method 1[a]	method 2[b]
subset 1	0.9444	0.9200	1.6399	2.3408
subset 2	0.9001	0.8144	0.9104	1.6290
training set	0.9367	0.8995	2.5503	3.9698
prediction set	0.8804	0.8725	2.6470	2.8556

[a] Method 1: PHMPSO when K takes two. [b] Method 2: QSAR study by modeling as a whole data set.

3.3. Nonpeptide Angiotensin II Antagonists Data. A set of eighty-five 1,2,4-triazoles[21] as nonpeptide angiotensin II antagonists, which were synthesized and evaluated for their antagonism against angiotensin II by Ashton et al., was used as a data set to further check the validity of the proposed methods. The parent structures of the compounds are presented in Figure 3a and Figure 3b, respectively. Table 2 of the Supporting Information lists the detailed structures of the compounds. The antagonism against angiotensin is expressed as IC_{50}, the molar concentration of the compound causing 50% antagonism of angiotensin II antagonists. We randomly divided the data set into two groups, a training set of sixty-seven compounds, and a prediction set of eighteen compounds. A series of molecular descriptors were 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, including the principal moment of inertia (PMI), the hydrophobic character (AlogP: logarithm of the partition coefficient in octano/water),[20] molar refractivity (MolRef), lowest unoccupied molecular orbital energy (LUMO) and electrotopological-state indices[22−24] (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 the atom; 's' stands for the single bond of the group; and 'OH' represents the hydroxyl radical.

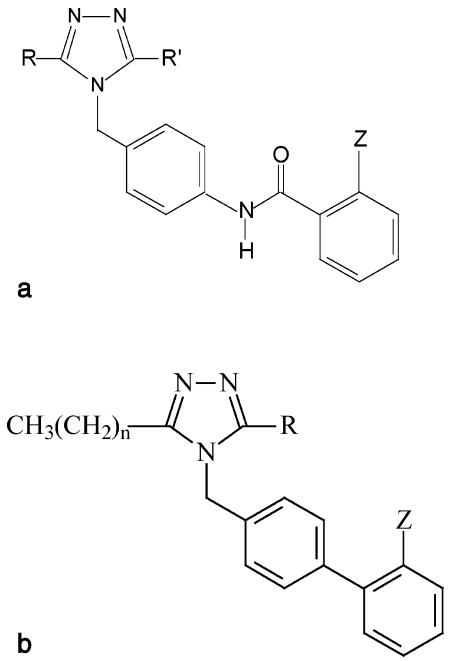

Figure 3. a. The parent structure of compound nos. 1−46 in Table 2 of the Supporting Information. b. The parent structure of compound nos. 47−85 in Table 2 of the Supporting Information.

Table 2. Results of QSAR Analysis of Inhibitors of Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinase by PHMPSO When the Whole Data Were Split into Two Subgroups Compared with that Obtained by MLR Modeling Using a Whole Data Set

	R (correlation coefficient)		RSS (sum of squared residual)	
data set	method 1[a]	method 2[b]	method 1[a]	method 2[b]
subset 1	0.8154	0.5322	4.9113	12.6253
subset 2	0.8842	0.8581	14.4901	18.1956
training set	0.8795	0.8002	19.4014	30.8209
prediction set	0.7915	0.6900	14.9445	20.7029

[a] Method 1: PHMPSO when K takes two. [b] Method 2: QSAR study by modeling as a whole data set.

The PHMPSO algorithm was 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. Piecewise Hypersphere Modeling by Particle Swarm Optimization for Hansch Data Set. The PHMPSO algorithm was first applied to the Hansch data set. Table 1 shows the results of this analysis. A correlation coefficient (R) of 0.9367 is obtained for compounds in the training set, as compared to the R value of 0.9030 in ref 17 and 0.8870 in ref 18. The sum of squared residuals (RSS) is 2.5503. The results were also compared by those obtained by MLR. As can be seen from Table 1 the PHMPSO algorithm improves the R and RSS of the model observably.

4.2. Piecewise Hypersphere Modeling by Particle Swarm Optimization for Data Set 1. A data set of sixty-one inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase with the corresponding inhibitory activities was taken to evaluate the PHMPSO algorithm. MLR analysis was also performed on this data set for a comparison. The top 5 best descriptors reported in the reference were used in the MLR model. The correlation coefficient (R) for the training set and the prediction set were 0.8002 and 0.6900, respectively. The correlation between the calculated and the experimental values of inhibitory activities is shown in Figure 4a. The squared residuals for the whole training data set are presented in Figure 4b. The sum of squared residuals (RSS) was 30.8209. As shown in Figure 4a and 4b, the correlation was rather poor and the modeling error is quite high. It seems difficult to construct one universal model for the whole population of interest with the desired residual.

For improving the QSAR modeling of inhibitors, PHMPSO was used to split the data set into submodels with improved linearity in each model. The same five descriptors were used in piecewise hypersphere modeling. In PHMPSO, hyperspheres are used for clustering all compounds in a training set to construct submodels. Particle swarm optimization is applied to search the optimal hyperspheres for finding satisfactory piecewise models. The population size of PSO is selected as 100. For this data set only two submodels are available as the further increase in the number of submodels violates the "five compounds per variable" rule. The result compared to the single model is summarized in Table 2. RSS is reduced from 30.8209 to 19.4014 for all compounds in the training set by PHMPSO. R in each submodel is higher than that in MLR modeling as a whole data set. The correlation between the calculated and experimental values of inhibitory activities is shown in Figure 4c. The squared residuals for the all observations are revealed in Figure 4d. The corresponding residuals by PHMPSO were obviously smaller than that in a single model by MLR. The prediction of compounds in the test set used the submodel selected by probability distance. A comparison of Figure 4a and Figure 4c and Figure 4b and Figure 4d shows that better results are obtained from multiple models by the PHMPSO algorithm than by MLR as a whole data set. The convergence process can be examined in Figure 5. As can be seen from Figure 5, the PHMPSO can converge to a satisfactory solution quickly. The time required to perform the algorithm is only several minutes. In one hypersphere subset it is common that the amino group of the 4-phenylamino moiety of the compounds participates in the conjugation of the quinazoline ring, while in the other hypersphere subset the amino group does not donate its electrons in the conjugation of the quinazoline ring. The presence of the 6,7-di-OMe groups is conducive to the activity. X-substituents at the 3-position of the 4-phenylamino moiety and Y-substituents at the 7-position enhance the activity. The increased electron density on the 1,3-N− in the quinazoline ring improves the binding of the molecules to the receptor.

4.3. Piecewise Hypersphere Modeling by Particle Swarm Optimization for Data Set 2. To further check the validity of the proposed methods, we applied the piecewise hypersphere modeling by particle swarm optimization for data set 2 to predict the antagonism of 85 1,2,4-triazoles, and the results were also compared to those obtained by MLR modeling. Five descriptors obtained by the variable selection analysis during GAs search were used to improve the MLR model. The correlation between the calculated and experimental values of antagonistic activities by MLR is disclosed in Figure 6a. Figure 6b reveals the squared residuals for the all observations by MLR modeling. The correlation coefficient of 0.7902 and 0.7253 are obtained for the training

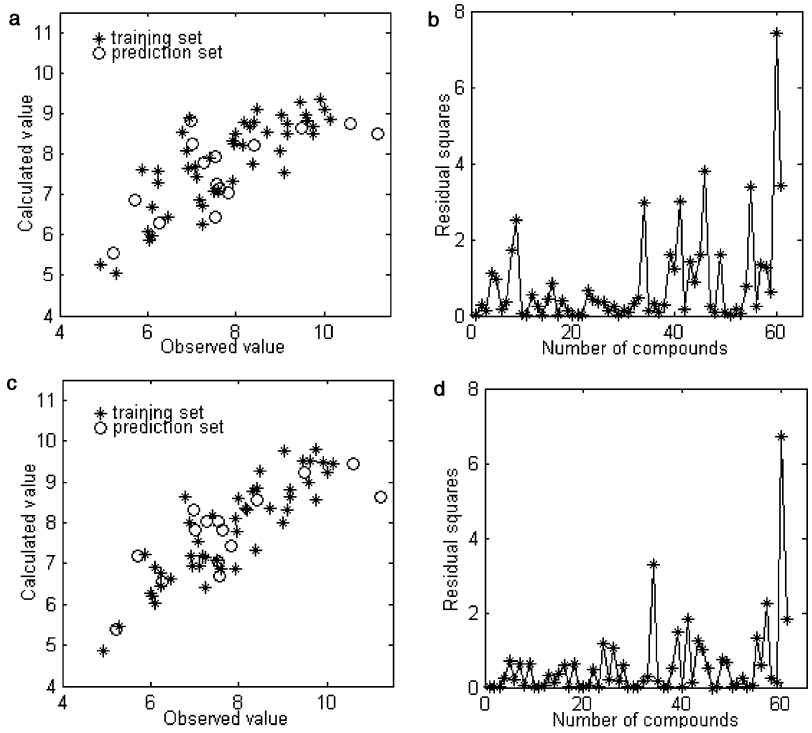

Figure 4. a. Calculated versus observed log 1/C of inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase by MLR modeling using a whole data set. b. The residual squares of inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase by MLR modeling using a whole data set. c. Calculated versus observed log 1/C of inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase by PHMPSO when the whole data was split into two subgroups. d. The residual squares of inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase by PHMPSO when the whole data was split into two subgroups.

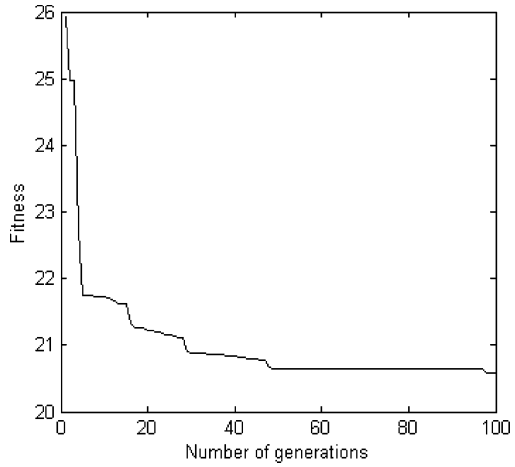

Figure 5. Convergence curve of inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase for PHMPSO when the whole data was split into two subgroups.

and prediction set, respectively. The proposed algorithm split the data set into 2 hypersphere subsets first. RSS is reduced from 17.8612 to 12.2004 for all compounds in the training set by the PHMPSO. The result of PHMPSO compared to the single model is presented in Table 3. The correlation between the calculated and experimental values of antagonistic activities is shown in Figure 6c. The squared residuals for the all observations are shown in Figure 6d. Using the same 5 variables, the corresponding residuals by PHMPSO were markedly smaller than that in a single model by MLR. The convergence curve of nonpeptide angiotensin II antagonists for the PHMPSO when the whole data was split into two subgroups is similar to that in Figure 5 (not shown). The results have revealed that using the proposed piecewise modeling to predict antagonisms of nonpeptide angiotensin II antagonists is likewise better than using a single model by MLR.

We also tried to split data set 2 to 3 hypersphere submodels by the PHMPSO. The result compared to the single model is presented in Table 4. The RSS is reduced from 17.8612 to 8.0837 for all compounds in the training set by the PHMPSO. The correlation between the calculated and experimental values of antagonistic activities is presented in Figure 6e. Compared to a single model by MLR, the correlation coefficient for the training set and the prediction set were improved from 0.7902 to 0.9110 and from 0.7253 to 0.8392, respectively. The squared residuals for all the observations are revealed in Figure 6f. One may notice that the RSS of compounds in the prediction set modeled by the PHMPSO is even larger compared to the RSS of these compounds in a single model by MLR. As can be seen from Figure 6e, one sample which did not belong to any hypersphere subset actually was predicted by an incorrect subset model, thus its squared residual was significantly higher than that by MLR. Owing to this, the RSS of compounds in the prediction set became augmented. As R is a coefficient to weigh the unitary correlativity for the calculated and experimental values of antagonistic activities, the outlier has little influence on R. The RSS of compounds in the training set was reduced greatly compared with that when K was 2. It demonstrates that for this data set it will be more suitable to split the whole data set into 3 hypersphere subsets. In the first hypersphere subset it is found that the values of AlogP and MolRef descriptors of the compounds in this subset tend to be smaller than those of the other two subsets. In the second hypersphere subset it is common that the values of descriptor LUMO are larger than the other two

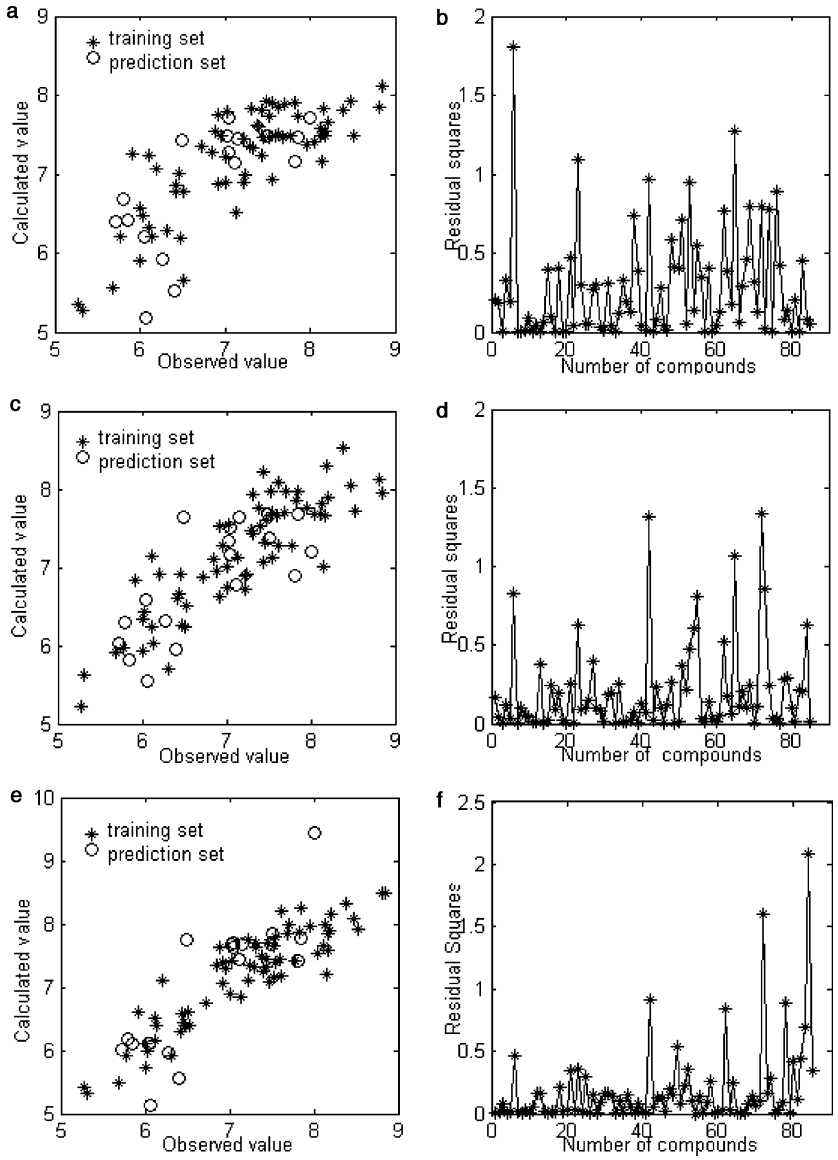

Figure 6. a. Calculated versus observed log(1/IC_{50}) of nonpeptide angiotensin II antagonists by MLR modeling using a whole data set. b. The residual squares of nonpeptide angiotensin II antagonists by MLR modeling using a whole data set. c. Calculated versus observed log(1/IC_{50}) of nonpeptide angiotensin II antagonists by PHMPSO when the whole data was split into two subgroups. d. The residual squares of nonpeptide angiotensin II antagonists by PHMPSO when the whole data was split into two subgroups. e. Calculated versus observed log(1/IC_{50}) of nonpeptide angiotensin II antagonists by PHMPSO when the whole data was split into three subgroups. f. The residual squares of nonpeptide angiotensin II antagonists by PHMPSO when the whole data was split into three subgroups.

Table 3. Results of QSAR Analysis of Nonpeptide Angiotensin II Antagonists by PHMPSO When the Whole Data Were Split into Two Subgroups Compared with that Obtained by MLR Modeling Using a Whole Data Set

	R (correlation coefficient)		RSS (sum of squared residual)	
data set	method 1[a]	method 2[b]	method 1[a]	method 2[b]
subset 1	0.6430	0.5495	9.1370	10.9753
subset 2	0.9365	0.8512	3.0634	6.8858
training set	0.8622	0.7902	12.2004	17.8612
prediction set	0.7468	0.7253	4.7084	5.6424

[a] Method 1: PHMPSO when K takes two. [b] Method 2: QSAR study by modeling as a whole data set.

Table 4. Results of QSAR Analysis of Nonpeptide Angiotensin II Antagonists by PHMPSO When the Whole Data Were Split into Three Subgroups Compared with that Obtained by MLR Modeling Using a Whole Data Set

	R (correlation coefficient)		RSS (sum of squared residual)	
data set	method 1[a]	method 2[b]	method 1[a]	method 2[b]
subset 1	0.9360	0.8724	2.3281	4.9623
subset 2	0.9599	0.4617	0.2672	3.7346
subset 3	0.7366	0.5112	5.4884	9.1643
training set	0.9110	0.7902	8.0837	17.8612
prediction set	0.8392	0.7253	7.5454	5.6424

[a] Method 1: PHMPSO when K takes three. [b] Method 2: QSAR study by modeling as a whole data set.

subsets. The LUMO descriptor, which measures the electrophilicity of a molecule, is important in governing molecular reactivity and properties. Molecules with low-lying LUMOs are more able to accept electrons than those with high-energy LUMOs. AlogP, the octanol/water partition coefficient, and molar refractivity are molecular descriptors relating chemical structure to observed chemical behavior. AlogP is related to the hydrophobic character of the molecule. The molecular

refractivity index of a substituent is a combined measure of its size and polarizability. The substituent variations in hydrophobic, electronic, and steric properties have important effects on the activities of the compounds, which leads to the different allocations of the compounds. This piecewise model shows satisfactory prediction performance in the QSAR analysis of antagonism of angiotensis II antagonists even when there are outliers existing.

4.4. Selection of Parameters for PHMPSO. The number of hyperspheres, K, i.e., the number of submodels, is determined ultimately in view of the two factors: if the residuals of the model reduce remarkably with the number of submodels (K) increasing; meanwhile, the number of compounds in each submodel compared to the number of variables ought to obey the rule of thumb.

The parameter w is an inertia weight which is brought into eq 1 to balance the global search and local search. By experience w is set to 0.5.

The time parameter μ in eq 2 is a random number uniformly distributed in [0,1]. It determines the different flying time for each particle.

The parameter ρ in the fitness function (eq 3) is the weighting coefficient between the size of the residuals from linear fitting within each subset and the compactness of clusters in the data space. The larger the value of the parameter ρ, the more compact the clusters. But this will cause the residuals from linear fitting within each subset to increase. Accordingly, ρ is set to 0.1 by experience to keep balance between the residuals from linear fitting within each subset and the compactness of clusters in the data space.

5. CONCLUSION

This paper developed piecewise hypersphere modeling by particle swarm optimization (PHMPSO) and introduced a new objective function to determine the appropriate piecewise models. Hyperspheres are used for clustering all compounds in a training set to construct submodels, and PSO is applied to search the optimal hyperspheres for finding satisfactory piecewise linear models. The PHMPSO algorithm has been first tested using the Hansch data set to examine its ability. Inhibitory activities of inhibitors of the epidermal growth factor receptor (EGFR) tyrosine kinase and antagonisms of nonpeptide angiotensin II antagonists were predicted by the proposed piecewise modeling. Experimental results have shown the good performance of this technique in improving the QSAR modeling. The PHMPSO algorithm has been designed for improvement of the performance of regression models when the data set contains a structurally diverse set of compounds. If such a structural diversity does not exist, such a methodology would not be necessary, though the method would automatically cluster the whole data set into only one group and perform the regression in a conventional way.

ACKNOWLEDGMENT

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

Supporting Information Available: Detailed structural formulas of the compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Tetko, I. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 2. Variable Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794−803.
(2) Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GA-Based PLS Analysis of Calcium Channel Antagonists. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306−310.
(3) 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*, 112−120.
(4) Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure−Property Relationship Approach Based on the *k*-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.
(5) Hoskuldsson, A. Variable and Subset Selection in PLS Regression. *Chemom. Intell. Lab. Syst.* **2001**, *55*, 23−38.
(6) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218−1227.
(7) Xiao, Z.; Xiao, Y.-D.; Feng, J.; Golbraikh, A.; Tropsha, A.; Lee, K.-H. Antitumor Agents. 213. †Modeling of Epipodophyllotoxin Derivatives Using Variable Selection *k* Nearest Neighbor QSAR Method. *J. Med. Chem*. **2002**, *45*, 2294−2309.
(8) Liu, S.-S.; Liu, H.-L.; Yin, C.-S.; Wang, L.-S. VSMP: A Novel Variable Selection and Modeling Method Based on the Prediction. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 964−969.
(9) Cho, S. J.; Hermsmeier, M. A. Genetic Algorithm Guided Selection: Variable Selection and Subset Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 927−936.
(10) Kennedy, J.; Eberhart, R. Particle Swarm Optimization. *In IEEE Int'1 Conf on Neural Networks* **1995**, *4*, 1942−1948.
(11) Shi, Y.; Eberhart, R. A modified particle swarm optimizer. *In IEEE World Congress on Computational Intelligence* **1998**, 69−73.
(12) Clerc, M.; Kennedy, J. The particle swarm-explosion, stability and convergence in a multidimensional complex space. *IEEE Transactions on evolutionary computation* **2002**, *6*, 58−64.
(13) Shi, Y.; Eberhart, R. Fuzzy adaptive particle swarm optimization. *In Proc Congress on Evolutionary Computation* **2001**, 101−106.
(14) Kennedy, J.; Eberhart, R. A discrete binary version of the particle swarm algorithm. *IEEE Int'1 Conf on Computational Cybernetics and Simulation* **1997**, 4104−4108.
(15) Shen, Q.; Jiang, J. H.; Jiao, C. X.; Huan, S. Y.; Shen, G. L.; Yu, R. Q.Optimized partition of minimum spanning tree for piecewise modeling by particle swarm algorithm. QSAR studies of antagonism of angiotensin II antagonists. *J. Chem. Inf. Comput. Sci.* in press.
(16) Shen, Q.; Jiang, J. H.; Jiao, C. X.; Lin, W. Q.; Shen, G. L.; Yu, R. Q. Hybridized particle swarm algorithm for adaptive structure training of multilayer feed-forward neural network: QSAR studies of bioactivity of organic compounds. *J. Comput. Chem.* **2004**, *25*, 1726−1735.
(17) Hansch, C.; Silipo, C.; Steller, E. E. Formulation of de novo substituent constants in correlation analysis: inhibition of dihydrofolate reductase by 2,4-diamino-5-(3,4-dichlorophenyl)-6-substituted pyrimidines. *J. Pharm. Sci.* **1975**, *64*, 1186−1191.
(18) Kubinyi, H. Evolutionary variable selection in regression and PLS analysis. *J. Chemom*. **1996**, *10*, 119−133.
(19) Alka, K.; Rajni, G.; Corwin, H. Comparative QSAR Study of Tyrosine Kinase Inhibitors. *Chem. Rev.* **2001**, *101*, 2573−2600.
(20) 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.
(21) Ashton, W. T.; Cantone, C. L.; Chang, L. L. Nonpeptide angiotensin II antagonists derived from 4H-1,2,4-triazoles and 3H-imidazo trialzoles. *J. Med. Chem.* **1993**, *36*, 591−609.
(22) Kier, L. B.; Hall, L. H. Derivation and Significance of Valence Molecular Connectivity. *Pharm. Res.* **1990**, *7*, 801−807.
(23) 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*, 76−82.
(24) 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.

CI049642M