Artificial Intelligence in Medicine 49 (2010) 61–66

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Modified tabu search approach for variable selection in quantitative structure–activity relationship studies of toxicity of aromatic compounds

Qi Shen, Wei-Min Shi *, Wei Kong

Department of Chemistry, Zhengzhou University, Zhengzhou 450052, China

ARTICLE INFO

Article history:
Received 3 December 2007
Received in revised form 2 November 2009
Accepted 17 January 2010

Keywords:
Tabu search
Variable selection
Quantitative structure–activity relationship
Aromatic compound

ABSTRACT

Objective: Variable selection is a key step in developing a successful quantitative structure–activity relationships (QSAR) analysis system. Tabu search (TS) can be used for variable selection which employs a flexible memory system to avoid convergence to local minima. But the convergence speed of TS depends on the initial solution and is slow. It usually reaches local minima since a single candidate solution is used to generate offspring. In the present paper, the TS algorithm was modified to assist TS to find the promising regions of the search space rapidly.

Methods and materials: A version of modified TS algorithm is proposed to select variables in QSAR modeling and to predict toxicity of some aromatic compounds. In the modified TS, the information which shares mechanism among the best position of all iteration and the personal position is introduced in the step of generating neighbors of the given solution. The move function which directs the moving of the solution is recorded as tabu. The modified Cp statistic is employed as fitness function.

Results and conclusions: For comparison, the conventional TS and stepwise regression were also examined. Experimental results demonstrate that the modified TS is a useful tool for variable selection which converges quickly towards the optimal position.

1. Introduction

Quantitative structure–activity relationships (QSAR) represent an attempt to search for quantitative relationship between chemical structural or property descriptors and activities by developing a QSAR model. Activities used in QSAR include chemical measurements, biological activity, toxicity or bioavailability, which are taken as dependent variables in building a model. Chemical structure is represented by a variety of descriptors, which include parameters to account for hydrophobicity, topology, electronic properties, and steric effects. QSAR data are characterized by hundreds even thousands of structural descriptors on only a few compounds with the biological activity values available. This leads either to possible overfitting and curse-of-dimensionality or even to a complete failure in building a meaningful regression model. Most descriptors may not be relevant to the given activity and these descriptors may potentially degrade the predictive performance of QSAR analysis by masking the contribution of the relevant descriptors. The selection of descriptors that are really indicative of activity concerned is one of the key steps in QSAR studies. The benefit gained from variable selection in QSAR data analysis is not only the improved predictive performance of the analysis model, but also the biological interpretability of relationship between the descriptors and biological activity. A large numbers of descriptors also increases computational complexity and is time-consuming. Therefore, variables selection is a key step in developing a successful QSAR analysis system.

The goal of variable selection is to find an "optimal" subset of all descriptors that maximizes information contents. To avoid the exponential explosion of an exhaustive search, several methods have been designed to determine the variable space in a more efficient way [1,2]. For the variable selections, the classical stepwise regression procedure can be used, as well as some more sophisticated techniques such as simulated annealing (SA) [3], genetic algorithms (GAs) [4,5] and evolution algorithm (EAs)[6]. Among these, GAs and EAs are randomized search algorithms that attempt to overcome the computational costs of exponential methods and are classified as a category of the research of so-called artificial life. Tabu search (TS), a relatively new optimization technique in this category, can also be used as a powerful optimizer which has been successfully applied to a number of combinatorial optimization problems [7–15]. It employs a flexible memory system to avoid convergence to local minima. But the convergence speed of TS depends on the initial solution and is slow [16]. It usually reaches local minima since a single candidate solution is used to generate offspring [17]. In the present paper, the TS

* Corresponding author. Tel.: +86 371 67767957; fax: +86 371 67763220.
E-mail address: shiweimin@zzu.edu.cn (W.-M. Shi).

0933-3657/$ – see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.artmed.2010.01.004

algorithm was modified to assist the TS to find the promising regions of the search space rapidly. A modified TS algorithm was proposed to select variables in multiple linear regression (MLR) modeling and used to predict toxicity of aromatic compounds. The results were compared to those obtained by stepwise regression. The results demonstrate that the modified TS is a useful tool for variable selection which converges quickly towards the optimal position.

Aromatic compounds are widely used in papermaking, leather, textile and other industries. Benzene derivatives comprise a significant component of the pollutant burden on the environment and have known harmful effects on human health and the environment. Experimental assessment of toxicity of aromatic compounds can be expensive, time-consuming and hazardous. QSAR can be used to predict toxicity of aromatic compounds when experimental data are not available.

2. Methods

2.1. Tabu search

TS which is a metaheuristic strategy was initially proposed by Fred Glover [7]. Lots of applications of TS can be found in tutorials [7–9]. TS is an iterative procedure designed for the solution of optimization problems. TS starts with a random solution or a solution obtained by a constructive and deterministic method and evaluates the fitness function. Then all possible neighbors of the given solution are generated and evaluated. A neighbor is a solution which can be reached from the current solution by a simple, basic transformation or move. New solution is generated from the neighbors of the current one. To avoid retracing the used steps, the method records recent moves in a tabu list. The tabu list keeps track of previously explored solutions and forbids the search from returning to a previously visited solution. If the best of these neighbors is not in the tabu list, pick it to be the new current solution. One of the most important features of TS is that a new solution may be accepted even if the best neighbor solution is worse than the current one. In this way it is possible to overcome trapping in local minima. If a neighbor solution is selected as new solution, this solution or moves is classified as tabu. Some aspiration criteria which allow overriding of tabu status can be introduced if that moves is found to lead to a better fitness with respect to the fitness of the current optimum. The aspiration criterion selected here will avoid missing good solutions. If the best object function of the generation fulfills the end condition or the number of iteration reaches a user-defined limit, the algorithm stops. Otherwise, the algorithm continues the TS procedures. TS method usually is completed with diversification and intensification procedures.

For variable selection problem, TS was implemented as follows: (1) variable selection solution is represented by a 0/1 bit string. Initial solution is randomly generated. (2) The neighbor of a solution vector x is a set of solutions, which are generated through adding or deleting a variable on x. (3) Twenty neighbors solutions are randomly selected as candidate solutions. (4) If a neighbor solution is selected as the new current solution, this move is recorded in tabu list. The tabu list size is selected as 5. (5) If a neighbor solution results in a best fitness for all previous iterations, pick it to be the new current solution even if it is in the tabu list and record this move in tabu list. (6) Termination condition is a predefined number of iterations.

2.2. Modified TS

In TS algorithm, if the neighbor solution is not in tabu list, pick it to be the new current solution. However, this solution is often worse than the current best solution. On the other hand, it usually reaches local minima and the best solution in the TS is unchanged for a lot of iterations. It will take much time to reach the near-global minimum and the convergence speed of TS sometimes is slow [16,17]. To improve the performance, the information sharing mechanism among the best previous solution of all iteration and the current solution is introduced in the step of generating neighbors of a given solution. The neighbors are generated by moving the given solution following the best solution of all iteration and the move function directs the moving of it.

For D-variable selection in QSAR, the solution is expressed a string of binary bits in a D-dimensional space and is represented as $\mathbf{X} = (x_1, x_2, \ldots, x_D)$. The binary-bit-coded string stands for a set of variable, which are used for evaluating the fitness function (please refer to the next section). A bit "0" in a solution represents the uselessness of corresponding variable. A move in a search space is restricted to 0 or 1 on each dimension. The best previous solution that gives the best fitness value is represented as $\mathbf{P} = (p_1, p_2, \ldots, p_D)$. In binary problem, updating a solution represents changes of a bit which should be in either state 1 or 0. The move function ($\mathbf{F}$) which directs the moving of the solution is represented as $\mathbf{F}_i = (f_1, f_2, \ldots, f_D)$. The move function f_d of every neighbor ($\mathbf{N} = (n_1, n_2, \ldots, n_D)$) is a random number in the range of (0, 1). The resulting move in solution is then defined by the following rule:

$$\text{If } (0 < f_d \le a), \text{ then } n_d = x_d \tag{1}$$

$$\text{If } (a < f_d), \text{ then } n_d = p_d \tag{2}$$

where a is a random value in the range of (0, 1) named static probability. The static probability a plays the role of balancing the global and local search. That means the larger the value of the parameter a is, the greater the probability for the neighbors to overleap local optima. On the other hand, a small value of parameter a is favorable for the neighbors to follow the best past positions and for the algorithm to converge more quickly. Therefore, we define the parameter descending along with the generation. Static probability a starts with a value 0.7 and decreases to 0.3 when the iteration terminates. As introduction of the information sharing mechanism, the random solution tends to converge to the best solution quickly. The modified TS is described as follows:

- Step 1. Randomly generate an initial solution ($\mathbf{X}$, initial binary string) and evaluate the fitness function of this individual. The initial solution is a string of binary bits corresponding to each variable.
- Step 2. Generate the neighbors of the solution according to the information sharing mechanism (Eqs. (1) and (2)). Then the performance of each neighbor or solution is measured according to a pre-defined fitness function. When setting the number of neighbor solution as a large value, the algorithm would make a deep search in a local region. But this may increase the calculation burden. Hence, the number of neighbors solution is set as 20 for variable selection by experience to keep balance between the search depth and calculation burden.
- Step 3. Pick new individual from the examined neighbor according to the aspiration criteria and tabu conditions and then update the solution. The move function which directs the moving of the solution was recorded in tabu list. If the neighbor solution is not in tabu list, pick it to be the new current solution. If the best of these neighbors is found to lead to a better fitness with respect to the fitness of the current optimum, override the tabu status and pick it to be the new current solution according to aspiration criteria. A small value of the tabu list size is favorable to reduce the calculation burden. However setting the tabu list size as a small value may cause the algorithm to

converge to local optima. The tabu list size was selected as 5 to jump out from local minima.

- Step 4. If all neighbors are tabu solutions, a new solution is generated randomly to further improve the ability of modified TS to overleap local optima. The fitness function of the new solution is then evaluated.
- Step 5. If the number of iteration reaches a pre-defined number of iterations, the training stopped with the results output, otherwise, go to the second step to renew solution. The modified TS scheme is presented in Fig. 1.

2.3. Fitness function

In modified TS, the performance of each neighbor or solution is measured according to a pre-defined fitness function. The modified Cp statistic as objective function is applied to variable selection in the modified TS. The modified Cp in MLR is expressed as follows:

$$\mathrm{Cp}(p) = \frac{\mathrm{RSS}_p}{\hat{\sigma}^2_{\mathrm{PLS}}} - (n - 2p) \quad (3)$$

where n is the number of dependent variables and p is the number of independent variables. RSS_p is the residual sum of the squares of p-variable MLR model, $\hat{\sigma}^2_{\mathrm{PLS}}$ is defined as the value of RSS corresponding to the minimum number of principal components in conventional partial least squares (PLS) analysis of the original data set when further increase of the number of principal components does not cause a significant reduction in RSS. PLS analysis is a factor analytical technique that is useful when there are more independent variables in data matrix (X) than in the target matrix (Y). The underlying assumption of PLS is that the observed data is generated by a process which is driven by a small number of latent variables or principal components. In its general form PLS creates orthogonal score vectors or components by maximizing the covariance between different sets of variables. The details of modified Cp have been described elsewhere [18].

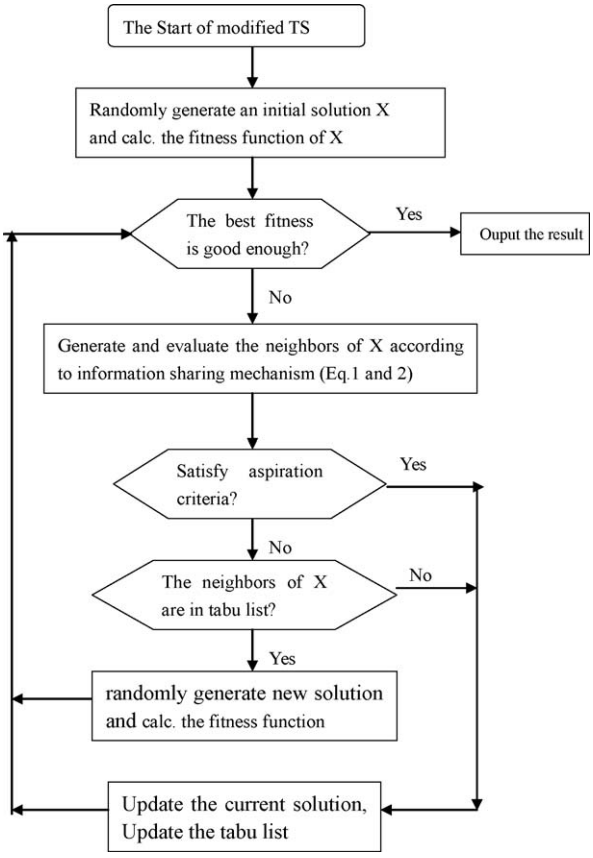

Fig. 1. The chart of the modified TS scheme.

3. Aromatic compounds toxicity data

A total of 65 aromatic chemicals with the observed toxicity to chlorella vulgaris in a novel short-term assay taken from the study by Netzeva et al. [19] were used to assess the performance of the modified TS in variable selection of QSAR. The data set is chemically heterogeneous (includes phenols, anilines, nitrobenzenes, benzaldehydes, *etc.*) and represents several mechanisms of toxic action.

A series of descriptors were calculated, which encoded different aspects of the molecular structure and consist of spatial, thermodynamic, structural, electronic and information-content descriptors. The spatial descriptors [20,21] used involve radius of gyration (RadOfGyration), density, molecular surface area, principal moment of inertia (PMI), molecular volume, and shadow indices. The thermodynamic descriptors [22] were taken to describe the hydrophobic character, refractivity (MolRef: molar refractivity), heat of formation (Hf) and the dissolution free energy for water and octanol (Fh2o: desolvation free energy for H_2O; Foct: desolvation free energy for octanol). Structural descriptors include the molecular weight (MW), the number of rotatable bonds (Rotbonds) and the number of hydrogen bond (Hbond acceptor). The electronic descriptors [23] taken were concerning superdelocalizability (Sr), atomic polarizabilities (Apol), and the dipole moment (Dipole). Electrotopological-state indices (E-State indices) [24,25] involved S-aasC, S-aaN, S-aNH2, *etc*. Some so-called information-content descriptors [26] such as atomic composition indices and multigraph information-content indices were also included in the candidate list. All these molecular descriptors were generated using Cerius2$^{3.5}$ software on Silicon Graphics R3000 workstation. Besides the aforementioned molecular descriptors, 7 variables used by Netzeva et al. were also included in the list of the candidate variables ($\log K_{\mathrm{ow}}$: logarithm of the octanol–water partition coefficient; E_{homo}: MOPAC energy of the highest occupied molecular orbital; E_{lumo}: energy of the lowest unoccupied molecular orbital; A_{max}: maximum acceptor superdelocalizability; Q_{Hmax}: maximum positive partial charge on a hydrogen atom; Q_{min}: maximum negative partial charge; ${}^0\chi^{\mathrm{v}}$: molecular connectivity indices).

The descriptor analysis involves the detection and removal of those structural descriptors which exhibit high pair-wise correlations with other descriptors or which contain little discriminatory information. Pairs of descriptors that are highly correlated ($r \geq 0.90$) encoded similar information, and one of them should be removed. Descriptors that contain a high percentage ($\geq$90%) of identical values are also discarded. Table 1 summarizes all molecular descriptors used as the candidate variables for selection.

The modified TS and MLR algorithms 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

The modified TS was first used for variable selection, in which the number of neighbors solution was set as 20 and the tabu list size was selected as 5. The number of iterations was set as 1000. The selected descriptors were taken as dependent variables to build QSAR models with MLR method. The best model with minimum fitness value contains three variables as given by the modified TS. The three variables are *Alog*P, MolRef and A_{max}. The correlation between the calculated and experimental values of lg1/EC_{50} of three-descriptor model is shown in Fig. 2. The correlation

Table 1
List of molecular descriptors for aromatic compounds studied as candidate variables.

Functional families of descriptors	Descriptors
Spatial descriptors	Shadow indices (surface area projections) (Shadow-XY, Shadow-XZ, Shadow-YZ, Shadow-nu, Shadow-XYfrac, Shadow-XZfrac, Shadow-YZfrac) Vm (molecular volume) Density Area (molecular surface area) RadOfGyration (Radius of gyration) Jurs descriptors (Jurs Charged Partial Surface Area descriptors) PMI (Principal moment of inertia, including PMI-mag, PMI-X, PMI-Y, PMI-Z)
Structural descriptors	MW (molecular weight) Hbond acceptor (number of hydrogen bond acceptors) Hbond donor (number of hydrogen bond donors) Rotbonds (number of rotatable bonds)
Electronic descriptors	Apol (sum of atomic polarizabilities) Dipole (Dipole-mag, Dipole-X, Dipole-Y, Dipole-Z) Sr (superdelocalizability)
Quantum mechanical descriptors	HOMO, E_{homo} (highest occupied molecular orbital energy) LUMO, E_{lumo} (lowest unoccupied molecular orbital energy) A_{max} (maximum acceptor superdelocalizability) Q_{max} (maximum positive partial charge on a hydrogen atom) Q_{min} (maximum negative partial charge)
Thermodynamic descriptors	*A*log*P*, log*P*, logK_{ow} (the octanol/water partition coefficient) Fh2o (desolvation free energy for water) Foct (desolvation free energy for octanol) MolRef (molar refractivity) Heat of formation (Hf)
E-State index	S-sCH3, S-aaCH, S-aasC, S-aNH2, S-aasN, S-ssO, S-sOH
Topological index	$^{0}\chi^{v}$, $^{1}\chi^{v}$, $^{2}\chi^{v}$, $^{3}\chi^{v}$, $^{4}\chi^{v}$, $^{0}\chi$, $^{1}\chi$, $^{2}\chi$, $^{3}\chi$, $^{4}\chi$

coefficient (R^2) and the standard deviation for the three-variable model is 0.8664 and 0.3941, respectively. The two-variable model obtained by modified TS is the same as those obtained by Netzeva in the literature [19]. These models and the best model that contain four variables are shown in Table 2. In these equations, positive coefficient of *A*log*P* indicates that the large hydrophobic character of the molecule would promote the toxicity. MolRef is a combined measure of molecular size and polarizability, and positive coefficient of MolRef indicates that increasing the molar refractivity of the molecule causes an increase in toxicity. These results shows that A_{max} is one of the important variables and the larger A_{max} can increase the toxicity. That is in accordance with the conclusions by Netzeva et al. that A_{max} was found to be a superior descriptor for modeling the acute toxicity of aromatic compounds [20].

Variable selection by conventional TS was also performed to compare with the modified TS. The best model with minimum fitness value given by the conventional TS contains two variables. The two variables are logK_{ow} and E_{lumo}. The correlation coefficients (R^2) and the standard deviation were 0.8389 and 0.4291, respectively. The minimum fitness could be obtained in about 40 cycles during the modified TS algorithm, but about 630 cycles were needed during conventional TS algorithm. Searching speed of the conventional TS algorithm is slow, but experimental results demonstrated that the modified TS converges to the global best solution rapidly.

To compare with modified TS, variables selection by stepwise regression was also performed. The obtained model by stepwise regression contains four variables. The four variables are Shadow-XYfrac, logK_{ow}, $^{0}\chi^{v}$ and A_{max}. The correlation coefficients (R^2) and the standard deviation were 0.8433 and 0.4303, respectively. A comparison with stepwise regression shows that better results were obtained from modified TS algorithm.

The real goal of developing QSAR is to predict the activity. To check the validity of the proposed methods, the data set of 65 aromatic chemicals was stochastically divided into two groups. Two-third compounds were used as the training set for developing regression models, while the remaining one-third compounds were used as the predicted dataset. The best MLR models selected by the training set contain three variables (*A*log*P*, MolRef and A_{max}). The plot of observed toxicity against that calculated by the three-variable model is shown in Fig. 3. It was found that using

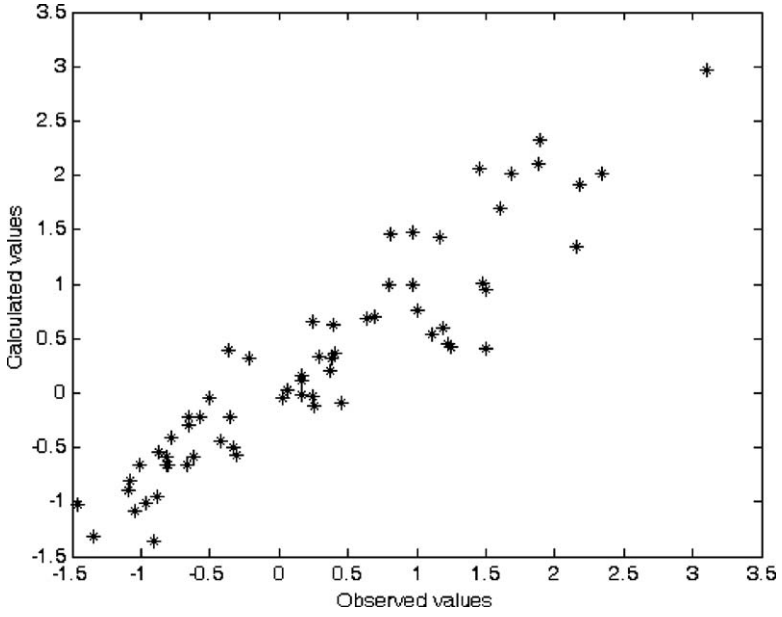

Fig. 2. Calculated versus observed lg1/EC_{50} of three-variable model by modified TS using multiple linear regression modeling.

Table 2
Results of variable selection by the modified TS and MLR modeling.

No.	Equation	R^{2a}	S^a	F^a
1	$\lg 1/IE_{50} = 0.731 \times \log K_{ow} - 0.5906 \times E_{lumo} - 1.9142$	0.8389	0.4291	161.4795
2	$\lg 1/IE_{50} = 0.4880 \times AlogP + 0.0314 \times MolRef + 19.4433 \times A_{max} - 8.6175$	0.8664	0.3941	131.7711
3	$\lg 1/IE_{50} = 0.4340 \times AlogP + 0.0375 \times MolRef + 18.0856 \times A_{max} - 0.0551S\text{-}ssO - 8.2768$	0.8729	0.3874	103.0543

[a] R: correlation coefficient; S: standard deviation; F: F-statistics.

three descriptors the correlation coefficients (R^2) for the training and the test set were 0.8709 and 0.8597, respectively. The standard deviation for training set was 0.3813.

Even the data set was stochastically divided into two groups, it should be noted that the predicted accuracy of a model at each iteration is not necessarily the same because of the various partition of training and tests sets. The reliability of a model is an essential issue in QSAR analysis. To evaluate the predictive ability and reliability of models by the modified TS accurately, the total samples were randomly partitioned into training and tests sets 200

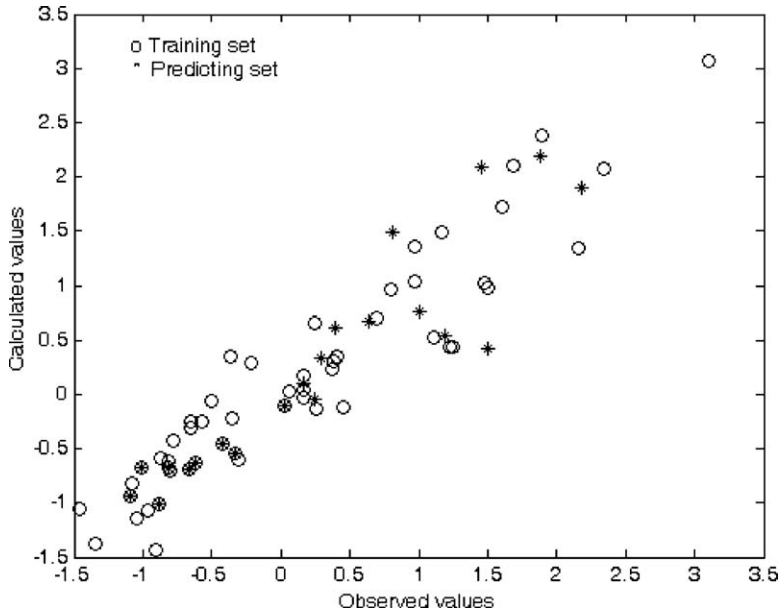

Fig. 3. Plot of $\lg 1/EC_{50}$ calculated from three-variable model versus the observed $\lg 1/EC_{50}$ values for training and test set.

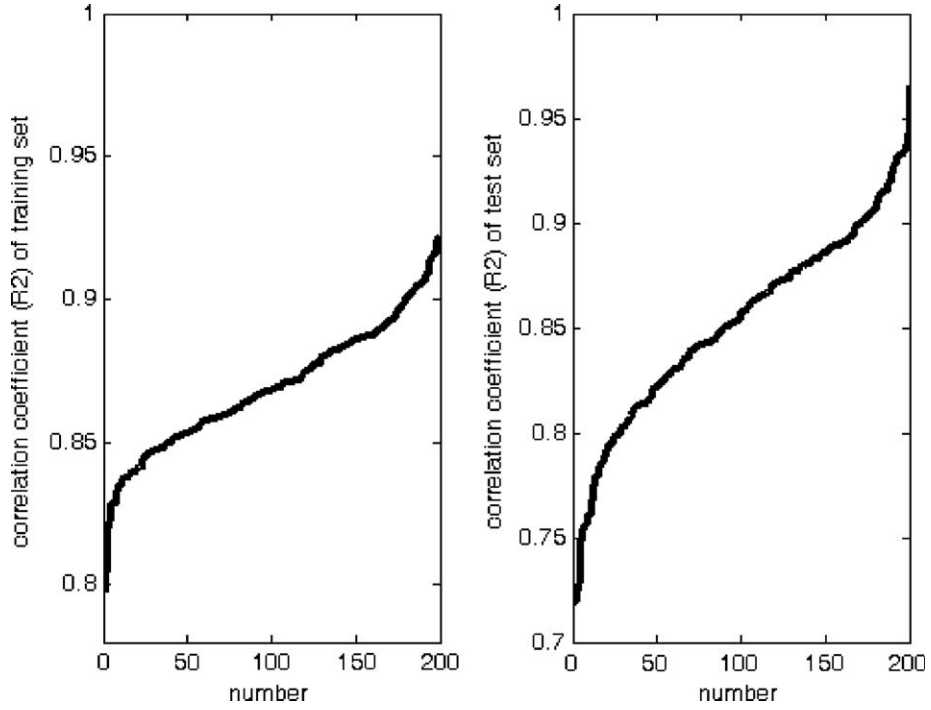

Fig. 4. Distribution of correlation coefficient (R^2) over 200 runs of partition samples using the best three-variable model.

times and then averaged the model accuracy for each partition of training and tests sets. By resampling a large number of learning samples, the correlation coefficient (R^2) for training set and test set were 0.8682 and 0.8515, respectively by the modified TS. The root mean square errors for training set and test set were 0.3582 and 0.4646, respectively. Fig. 4 shows the distribution of correlation coefficient (R^2) over 200 runs of partition samples using the best three-variable model. As shown in Fig. 4, correlation coefficient (R^2) larger than 0.8515 is about 107 times in 200 runs for test set and the highest R^2 achieve 0.9650. The results show that the model by the modified TS is stable and reliable.

The modified TS was run for about 1000 times, and the number of times which a particular molecular descriptor appears in 1000 cycles was counted. When the descriptors by the order of decreasing numbers of times of appearance were listed, the top descriptors or the most frequently appeared feathers were obtained. A_{max} is an index of reactivity in aromatic compounds that was used by Netzeva et al. [19], and it turned to be one of the most important variables. MolRef is a combined measure of its size and polarizability which seems essential with respect to toxicity of aromatic compounds. *A*log*P* which is a factor relating to the hydrophobic character of the molecule are factors is usually much considered in the development of QSAR in biochemistry. They are shown to be important in QSAR of aromatic compounds toxicity. "E_{lumo}" is also a preferred descriptors and the negative coefficient of descriptor E_{lumo} implied that molecules with low-energy E_{lumo} would promote the toxicity. There are two electrotopological-state descriptors (S-sCH3, S-aNH2) among the top descriptors and these indices seem to be information-rich in describing molecular structure. The subscript '–sCH3' refers to methyl and 'aNH2' refers to the amine in phenyl group. 'NH2' in phenyl group can increase toxicity and 'CH3' in aromatic compounds would decrease toxicity. Besides these variables, descriptors Shadow-nu, Vm, Apol, Dipole-Z, PMI-mag, PMI-Z and Q_{Hmax} also have advantage for toxicity of aromatic compounds. The toxicity of aromatic compounds is a complex one, which involves spatial, thermodynamic, electronic, and structural effects.

5. Conclusion

In the present study, the TS algorithm was modified to be used in variable selection in QSAR modeling for predicting toxicity of aromatic compounds. The modified Cp was employed as fitness function. It has been demonstrated that the modified TS is a useful tool for variable selection with nice performance and the ability to select preferred variables with satisfactory convergence rates. In the selected descriptors, A_{max}, MolRef and *A*log*P* are the most important descriptors in predicting toxicity of aromatic compounds.

Acknowledgement

The work was financially supported by the National Natural Science Foundation of China (Grant no. 20505015).

References

[1] Gualdrón O, Llobet E, Brezmes J, Vilanova X, Correi X. Fast variable selection for gas sensing applications, vol. 2. In: Sensors, Proceedings of IEEE; 2004. p. 892–5.
[2] Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Boston: Kluwer Academic Publishers; 1998. p. 37–51.
[3] Shen M, Arnaud L, Xiao Y, Alexander G, Harold K, Alexander T. Quantitative structure–activity relationship analysis of functionalized amino acid anticonvulsant agents using *k* nearest neighbor and simulated annealing PLS methods. J Med Chem 2002;45:2811–23.
[4] Cho SJ, Hermsmeier MA. Genetic algorithm guided selection: variable selection and subset selection. J Chem Inf Comput Sci 2002;42:927–36.
[5] Yasri A, Hartsough D. Toward an optimal procedure for variable selection and QSAR model building. J Chem Inf Comput Sci 2001;41:1218–27.
[6] Brian TL. Evolutionary programming applied to the development of quantitative structure–activity relationships and quantitative structure–property relationships. J Chem Inf Comput Sci 1994;34:1279–85.
[7] Glover, F. Tabu search. Part II. ORSA J Comput 1990;2(1):4–32.
[8] Glover F, Laguna M. Tabu search. Boston: Kluwer Academic Publishers; 1997.
[9] Glover F, Laguna M. Tabu search, handbook of applied optimization. In: Pardalos PM, Resende MGC, editors. Oxford: Oxford University Press; 2002. p. 194–208.
[10] Pacheco J, Casado S, Núñez L, Gómez O. Analysis of new variable selection methods for discriminant analysis. Comput Stat Data Anal 2006;51(3):1463–78.
[11] Todeschini R, Consonni V, Pavan M. A distance measure between models: a tool for similarity/diversity analysis of model populations. Chemometr Intell Lab Syst 2004;70:55–61.
[12] Chavali S, Lin B, Miller DC, Camarda KV. Environmentally-benign transition metal catalyst design using optimization techniques. Comput Chem Eng 2004;28:605–11.
[13] Vainio MJ, Johnson MS. McQSAR: a multiconformational quantitative structure–activity relationship engine driven by genetic algorithms. J Chem Inf Model 2005;45(6):1953–61.
[14] Gani R, Harper PM, Hostrup M. Automatic creation of missing groups through connectivity index for pure-component property prediction. Ind Eng Chem Res 2005;44(18):7262–9.
[15] Mills Jamie D, Olejnik Stephen F, Marcoulides George A. The tabu search procedure: an alternative to the variable selection methods. Multivariate Behav Res 2005;40(*3*):351–71.
[16] Pad JS, Lut ZM, Chu SC, Sun SH. Non-redundant VQ channel coding using modified tabu search approach with simulated annealing. In: Third international conference on knowledge-based intelligent information engineering systems. Adelaide, Australia: IEEE Press; 1999. p. 242–5.
[17] Kvasnicka V, Pospichal J. Fast evaluation of chemical distance by tabu search algorithm. J Chem Inf Comput Sci 1994;34(5):1109–12.
[18] Shen Q, Jing JH, Yu RQ. Variable selection by an evolution algorithm using modified Cp based on MLR and PLS modeling: QSAR studies of carcinogenicity of aromatic amines. Anal Bioanal Chem 2003;375:248–54.
[19] Netzeva TI, Dearden JC, Edwards R, Worgan ADP, Cronin MTD. QSAR analysis of the toxicity of aromatic compounds to chlorella vulgaris in a novel short-term assay. J Chem Inf Comput Sci 2004;44:258–65.
[20] Rohrbaugh RH, Jurs PC. Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. Anal Chim Acta 1987;199:99–109.
[21] Stanton DT, Jurs PC. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. Anal Chem 1990;62:2323–9.
[22] Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. 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;29:163–72.
[23] Ciuprina G, Loan D, Munteanu I. Use of intelligent-particle swarm optimization in electromagentics. IEEE Trans Magn 2002;38(2):1037–40.
[24] Hall LH, Kier LB. The electrotopological state: structure information at the atomic level for molecular graphs. J Chem Inf Comput Sci 1991;31:76–8.
[25] Hall LH, Kier LB. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 1995;35:1039–45.
[26] Boncher. Danailinformation theoretic indices for characterization of chemical structures. Chichester, UK: Research Studies Press; 1983. p. 249.