

# Applications of genetic algorithms on the structure–activity correlation study of a group of non-nucleoside HIV-1 inhibitors

T.J. Hou, J.M. Wang, X.J. Xu \*

*Department of Chemistry, Peking University Molecular Design Laboratory of Jiuyuan Gene Engineering, Peking University, Beijing 100871, China*

## Abstract

Genetic algorithms (GAs) have been proven to be very useful in data analysis and can be applied as a very powerful technique in quantitative structure–activity relationship (QSAR) analysis. QSAR based on GAs allows the construction of models competitive with or superior to standard methods; moreover, from the analysis of the calculation results, we may get very useful additional information which cannot be provided by other methods. We developed a QSAR program combining genetic algorithm with multiple linear regression and cross-validation. We use it in the QSAR analysis of 23 HIV-1 inhibitors pyrrolbenzothiazepinones (PBTP) and pyrrolbenzoxazepinones (PBP). A group of suitable QSAR models has been obtained. Using the best model we predicted the RT activities of some compounds whose RT experimental activities are unknown. Moreover, from the statistical analysis of the multiple models, we found that low lipophilicity at C-6, small compounds surface, high  $\pi$  electron density of the benzo fused ring and low dipole along the  $z$  axis were the most important factors that may influence the RT activities. These descriptors allow a physical explanation of hydrophobic interaction, electronic and steric effect contributing to HIV-1 inhibitory potency. © 1999 Elsevier Science B.V. All rights reserved.

*Keywords:* Genetic algorithms (GAs); Quantitative structure–activity relationship (QSAR); HIV-1 reverse transcription inhibitors; Pyrrolbenzothiazepinone (PBTP); Pyrrolbenzoxazepinone (PBP)

## Contents

1. Introduction . . . . .	304
2. Methods . . . . .	304
2.1 QSAR based on GAs . . . . .	304
2.2. Choosing appropriate individual data structure . . . . .	305
2.3. Choosing adequate fitness function . . . . .	305
3. Experiment . . . . .	306
3.1. Experimental data . . . . .	306
3.2. Calculation details . . . . .	307
3.3. Results and discussion . . . . .	307

\* Corresponding author

4. Conclusion . . . . .	310
Acknowledgements. . . . .	310
References . . . . .	310

---

## 1. Introduction

Quantitative structure–activity relationship (QSAR) modeling provides a rational basis for understanding mechanisms of biological performance and how to improve performance by altering chemical structure. Current QSAR methods are mainly limited by the structure of the data: the number of compounds with requisite behavior measures (e.g., biological activity) is usually small compared with the number of features which can be measured or calculated. One of the most important and difficult problems in traditional quantitative structure–activity relationship is how to choose the adequate features to build the regression models. Recently some published papers suggested that genetic algorithms may be useful in QSAR analysis [1], especially the features selection in obtaining the proper QSAR models. We have developed a QSAR program based on GAs [2]. It has been used in our QSAR study. In most cases, very good results can be obtained.

Pyrrorobenzothiazepinones (PBTP) and pyrrolbenzoxapinones (PBP) which belong to non-nucleoside reverse transcriptase could become a new kind of potent and selective drugs against HIV-1 reverse transcriptase (HIV-1 RT) [4]. They can inhibit HIV-1 reverse transcriptase (RT) enzyme in vitro to prevent HIV-1 cytopathogenicity in T4 lymphocytes without appreciable activity on HIV-2 cytopathic effects and against HBV as well as calf thymus DNA  $\alpha$ -polymerase. Until now, no quantitative structure–activity relationships (QSAR) analysis of PBTP and PBP has been reported in the literature. A correlation study is expected to provide insight into the anti-HIV-1 mechanism of PBTP and PBP and give some useful information that can help researchers design new candidates as potential drugs.

## 2. Methods

### 2.1. QSAR based on GAs

Many methods, including CART, PCA and PLS, develop a single regression model by incremental addition or deletion of basis functions. In contrast, QSAR based on genetic algorithms uses a population of many models and tests only the final, fully-constructed models.

The brief basic steps of QSAR based on GAs are involved.

(1) *Creation of the initial population.* According to the genetic algorithm, an individual should be represented as a linear string, which plays the role of the DNA for the individual, so we randomly choose a series of features as a string. The initial population are generated by randomly selecting some numbers of features from the training set. Then these models are scored according to their fitness score. We use an elite population to remain the best and different individuals.

(2) *Crossover operation.* Once all models in the population have been rated using the fitness score, the crossover operation is repeatedly performed. In the operation, two good models are probabilistically selected as ‘parents’ with the likelihood of being chosen inversely proportional to a model fitness score.

(3) *Mutation operation.* After crossover operation, mutation operation may randomly alter all individuals in the new population, the new model fitness is determined.

(4) *Comparison operation.* After the crossover and mutation operation, we compare the newly created population and the elite population. If there are some individuals in the newly created population are better than some individuals in the elite population, we copy these better individuals to the elite population. If the

total fitness of the elite population could not be improved, we can say that ‘convergence’ is achieved.

(5) Partial reinitialization. In some cases, the convergence is achieved too fast, the individuals in the population largely change to the same, in this case the elite population are very difficult to improve, that is to say, it is trapped in a local minimum. How to escape from it, we proposed a partial reinitialization procedure, after a several steps of crossover and mutation operations, reinitialize some worst individuals in the population. Normally, we randomly reproduce the 80% individuals in the population. From study, we find that this procedure is effective to escape from local minima. Generally, three to six reinitializations are enough to find all different QSAR models.

Upon completion from the elite population, we can get the highest fitness score models. For a population of 200 models, if the data set contains about 30 features, 300–500 cycles are usually sufficient to achieve ‘convergence’, if the data set has 40 features, 1000–1500 operations are usually enough. For a typical data set this process takes 10 min to 1 h for a PC (Pentium 150).

## 2.2. Choosing appropriate individual data structure

It is well known that genetic algorithm is very flexible, there may be many variations in traditional genetic algorithm. Especially, for the data structure of the individuals, many data structures have been developed, for example: invariable length strings, variable length strings, matrix structure, tree type structure, etc. How to choose an adequate individual data structure is very important. Usually, we use invariable integral number string data structure for individuals in our QSAR analysis, the integral number string comprised the information of selected features and the user-specified basic functions. We think that it is simple, direct and effective in most QSAR analysis. This kind of data structure for individuals is very appropriate. Another data structure-variable length integral number string data structure for individuals has never been used in other QSAR analysis [2], the length of the string for the individuals can change according to the fitness score after crossover operation, this data structure can be used to build QSAR

models with different terms. Because different model has different terms so it is difficult to choose suitable fitness function. In Ref. [1], Rogers and Hopfinger used Freidman’s ‘lack of fit’ (LOF) measure [3] as the fitness function to evaluate the individuals. But there will appear many other problems if we use LOF as fitness function, the first problem is that it lacks sufficient mathematical backgrounds, the model with the best LOF do not thoroughly provide the best QSAR model; the second problem is LOF must be controlled by the users. For a new-user, there may be some difficulties in choosing the magnitude of parameter  $d$ . Moreover, from our study, we found that the GAs using the integral number string data structure are more convenient, more precise, more time-saving than using variable length integral number string data structure in most traditional QSAR analysis. So in this QSAR analysis, we use invariable integral number string data structure for individuals.

## 2.3. Choosing adequate fitness function

The goal of QSAR analysis is to build the reliable QSAR models. The critical factor of reliability to the models derived from the QSAR based on GAs is choosing a good fitness function. If we want to build a linear regression model, the simplest fitness function is to choose the multiple linear regression coefficient. But we think only relying on the multiple linear regression coefficient is not enough, at least we should validate this model which not only has low error measure on the training set, but also can predict well. So in our program, we often define the fitness to be the product of  $\{r_k\}$  and  $\{r_c\}$ . The  $\{r_k\}$  is the linear regression coefficient of the model and the  $\{r_c\}$  is the leave-one-out cross-validation coefficient of the model which appropriately links the multiple linear regression and cross-validation together. But strictly to say, evaluating a model only by a simple fitness score is often not enough, fitness score only tell you that this model maybe relatively good, but not absolutely good, we should evaluate it from many aspects including various experiments. In practice, you can choose the appropriate fitness function according to your needs. In our QSAR analysis, we use the  $r_c \cdot r_k$  as fitness function.

### 3. Experiment

In this study, genetic algorithm was applied to the QSAR analysis of PBTP and PBP which belong to non-nucleoside reverse transcriptase [4]. Due to their high specificity and low level of toxicity, these non-nucleoside inhibitors are the potential anti-AIDS drugs. A correlation study is expected to provide insight into the anti HIV-1 mechanism of PBTP and PBP and give some useful information that can help researchers to design new candidates for potential drugs.

#### 3.1. Experimental data

Biological activities of 22 PBTP and PBP derivatives against the cytopathic effect HIV-1 have been

reported as the test drug concentration which results in a 50% survival of uninfected untreated control CEMM-SS cells, e.g., cytotoxicity of the test drug ( $IC_{50}$ ,  $\mu\text{M}$ ). The potency has been defined as  $\log(1/C)$  in the QSAR analysis, where  $C$  is the  $IC_{50}$  of the compound, and is used as the independent variable in the QSAR study. Compounds showed in Fig. 1 were modeled using Cerius2 package. The initial structures were firstly minimized using molecular mechanics by using universal force field. The terminal condition was RMS gradient smaller than 0.001 Kcal/ $\text{\AA}$  mol). Quantum-chemical features were calculated using AM1 method of MOPAC7 package. The keyword PRECISE was used to get more accurate results. Hydrophobic coefficient was calculated based on the Crippen's fragmentation method [5]. The data set contains 22 compounds, 24 features, and a set

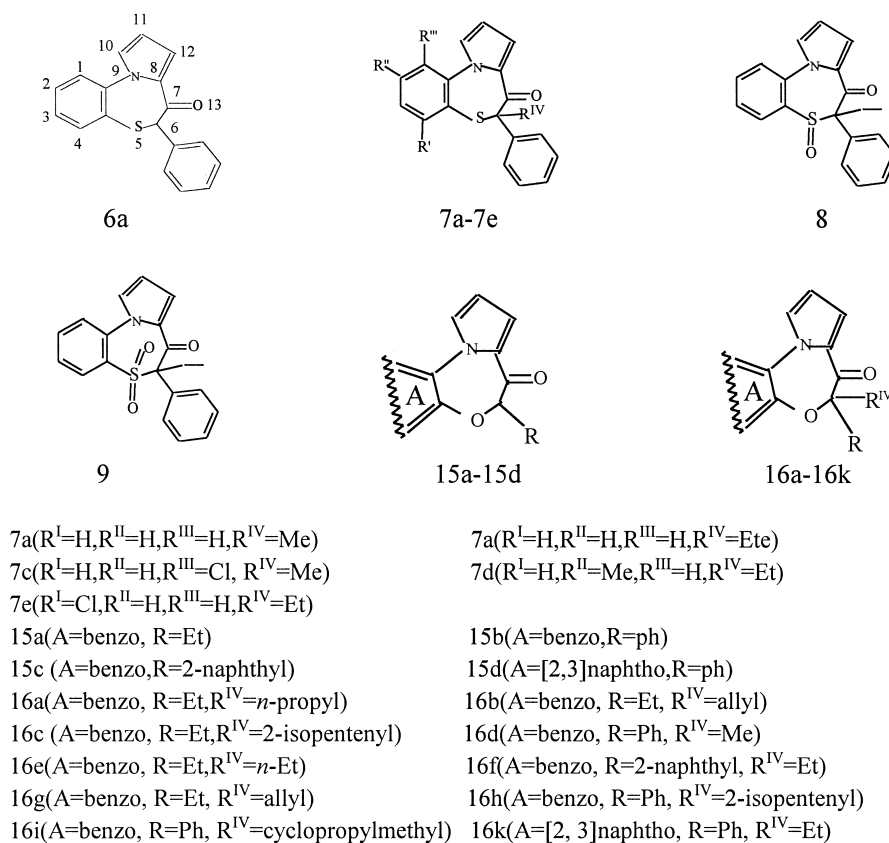


Fig. 1. 2D topographies of pyrrolobenzothiazepinones and pyrrolobenzoxazepinone.

- $\pi$ _mol: The hydrophobic coefficient of the molecules
- $\pi$ _frg : The hydrophobic coefficient of the substitutes in site 6
-surf : The surface area of the molecules
-volum: The connolly volume of the molecules
-Sfrg : The surface area of the substitutes in site 6
-E_for : The final heat of formation of the molecules
-E_ion: The ionization potential of the molecules
-weight: The molecular weight of the molecules
-E_home,E_lumo: The energy of home and lumo orbitors of the molecules
-Dip_x,Dip_y,Dip_z,Dip: The dipole vector and dipole vector components in x,y,z
-den_5,6,7,9: The atom electron density in site 5,6,7,9
-char_5,6,7,9: The atom charge derived from the electronic potential in cite 5,6,7,9
-den_pi1: The $\pi$ electron density of the benzo-fused ring
-den_pi2: The $\pi$ electron density of the conjugated on atoms 7,8,9,10,11,12,13

Fig. 2. The features that were used in the QSAR analysis in the data set.

of corresponding RT activities (all features and their abbreviation are listed in Fig. 2). Because some compound activities are not definite, so we only selected 15 compounds as training set. The data set was of particular interest because it contains a large number of features relative to the number of compounds.

### 3.2. Calculation details

We respectively selected four features and three features to search their best models. QSAR analysis based on GAs began with randomly generating a population of random models. These models were generated by randomly selecting three or four features from the data file. Product of multiple linear regression coefficient and leave-one-out cross-validation coefficient was used as fitness function to generate the fitness scores of these models. For this data set 200 populations were used, and the number of elite populations were 100. The genetic operator was applied until the total fitness score of the elite populations could not be improved over a period of 30 crossover operation. Moreover after 100 crossover and mutation operations, we applied a partial reinitialization procedure. The convergence criteria was met after 430 operations for four features and 280 operations for three features. The evolution took approximately 20 min for four features, about 10 min for three features. After convergence, we respectively got the 100 best models for four features and three features.

### 3.3. Results and discussion

After analysis, we got a large amount of satisfactory statistical models, the top 10 models are listed in Table 1. Results show that GA is very powerful to find the best models. We studied this quantitative structure–activity relationships of this group of compounds and proposed a QSAR model [6]. In our previous study, we used stepwise regression analysis method to model a QSAR model from 16 molecular parameters. From the results of the QSAR analysis with GAs, the model which we proposed in our previous study was discovered and was rated 72 out of 200. We can see that QSAR based on GAs can efficiently find a group of best models, so it afforded us more opportunity to select the best model that we require than other methods. From the fitness score, model 1 is the best, we can use it to predict the 7 compounds whose RT activities unknown. Table 2 lists all the four parameters and calculated activity data for each compound. The predicting RT activities of these compounds without RT experimental activities are listed in Table 2. The RT activity of compound 15d is the lowest.

If we use traditional methods, we can only get a single model. But sometimes a model with high linear regression coefficient sometimes does not mean it is absolutely a good model because some other factors may effect the result, e.g., random correlation, auto-correlation. In most cases the interaction between molecular features are very complex, as inter-

Table 1

Top 10 models generated using the training set

(1) $\log(1/C) = 4.916 - 0.906 * \text{Dip}_z + 2.574 * \pi_{\text{frg}} + 0.560 * \text{Dip}_y - 0.052 * \text{volum}$ Fitness = 0.8858 $F = 16.838$ ; $s = 0.486$ $r_k = 0.933$ ; $r_c = 0.892$	(2) $\log(1/C) = 6.481 + 3.088 * \text{char}_7 - 0.058 * \text{surf} - 0.396 * \text{Dip}_z + 0.060 * \text{Sfrg}$ Fitness = 0.8268 $F = 14.000$ ; $s = 0.502$ $r_k = 0.928$ ; $r_c = 0.891$
(3) $\log(1/C) = 17.751 + 5.230 * \text{cha}_n - 0.058 * \text{volum} - 1.414 * \text{E}_{\text{ion}} + 0.072 * \text{Sfrg}$ Fitness = 0.8146 $F = 13.922$ ; $s = 0.527$ $r_k = 0.921$ ; $r_c = 0.884$	(4) $\log(1/C) = 17.737 + 5.229 * \text{char}_n - 0.057 * \text{volum} + 1.413 * \text{E}_{\text{home}} + 0.072 * \text{Sfrg}$ Fitness = 0.8146 $F = 13.908$ ; $s = 0.527$ $r_k = 0.921$ ; $r_c = 0.884$
(5) $\log(1/C) = 183.281 - 4.901 * \text{den}_7 + 2.171 * \pi_{\text{frg}} - 0.579 * \text{Dip}_z - 0.054 * \text{surf}$ Fitness = 0.8142 $F = 15.805$ ; $s = 0.499$ $r_k = 0.921$ ; $r_c = 0.884$	(6) $\log(1/C) = 8.167 - 0.670 * \text{Dip}_z + 1.857 * \pi_{\text{frg}} - 0.66 * \text{surf} + 0.020 * \text{Sfrg}$ Fitness = 0.8114 $F = 14.000$ ; $s = 0.526$ $r_k = 0.921$ ; $r_c = 0.881$
(7) $\log(1/C) = 283.003 - 54.963 * \text{den}_9 + 0.057 * \text{Sfrg} - 0.340 * \text{Dip}_z - 0.050 * \text{surf}$ Fitness = 0.8106 $F = 13.34971$ ; $s = 0.537$ $r_k = 0.918$ ; $r_c = 0.883$	(8) $\log(1/C) = 0.775 + 0.724 * \text{E}_{\text{ion}} + 2.346 * \pi_{\text{frg}} - 0.761 * \text{Dip}_z - 0.058 * \text{surf}$ Fitness = 0.8050 $F = 13.922$ ; $s = 0.522$ $a_k = 0.92$ ; $a_c = 0.874$
(9) $\log(1/C) = 8.163 - 0.722 * \text{Dip}_z + 2.467 * \pi_{\text{frg}} - 0.063 * \text{surf}$ Fitness = 0.8043 $F = 18.756$ ; $s = 0.521$ $r_k = 0.915$ ; $r_c = 0.879$	(10) $\log(1/C) = 26.130 + 5.260 * \text{char}_9 - 0.072 * \text{surf} + 1.951 * \text{E}_{\text{homo}} + 0.068 * \text{Sfrg}$ Fitness = 0.8041 $F = 13.824$ ; $s = 0.529$ $r_k = 0.920$ ; $r_c = 0.874$

action of several features may result in another feature, moreover, only from a single good model, we may not get the most original factors influencing biological activity. So we think that we may get some useful information from multiple models rather than a single model. In our QSAR study, the data set is very small, so a single model is maybe not very reliable, we can average the results and get the averaging results. By averaging, the effect of models which are extrapolating beyond their predictive region may be reduced, so we can say sometimes that we may get more useful and deeper information from averaging the results of multiple models than an individual model. From the elite populations, we can select 73 models whose fitness scores are higher than 1.7, we count the features that appeared in these models. The statistical results are listed in Table 3.

There are totally 21 features appear in 73 best models in Table 3, but their appearing frequencies are quite different. The  $\pi_{\text{frg}}$ 's frequency in the models

is the highest, that is to say, this factor with many other features can generate good model. It is maybe the most important factor that affect the RT activity. Besides this factor, the frequency of surf, dip<sub>z</sub> and den<sub>pil</sub> is relatively large, these three factors are also maybe very important. Four factors and their parameters are listed in Table 4.

From the results of multiple model statistical analysis, we can see that the substituents at position 6 are very significant to RT activities, small lipophilic substituents at C-6 were preferred. Beside this factor, there are two electronic features are very important, high  $\pi$  electron density of the benzo-fused ring and low dipole along  $z$  can enhance RT activities, these two factors actually are affected by the electronic contribution on system. The  $\pi$  electron density is mainly influenced by substituents on the benzo-fused ring. The electron withdraw groups on benzo-fused ring may contribute to the compound RT activities. Moreover, the molecular surface area is very signifi-

Table 2

List of structural parameters used in model 1 and experimental and calculated RT activities

Comp_num	Dip_z	Dip_y	$\pi_{\text{frg}}$	Volum	$\log(1/C)$ (experimental)	$\log(1/C)$ (calculated)	Residue
Comp_6a	-0.889	1.235	2.03	250.88	-2.00	-1.60	0.40
Comp_7a	-0.364	1.355	3.12	273.70	-0.70	-0.38	-0.32
Comp_7b	-0.169	1.414	3.36	291.91	-1.17	-0.87	-0.30
Comp_7c	1.072	2.023	3.36	313.23	-2.78	-2.77	-0.01
Comp_7e	-0.269	-0.400	3.36	310.68	-2.70	-2.78	0.08
Comp_8	2.523	3.289	3.36	298.92	-2.70	-2.63	-0.07
Comp_15a	0.690	1.322	1.33	207.93	-2.00	-2.48	0.48
Comp_16a	0.200	1.281	2.66	243.91	-1.00	-0.55	-0.45
Comp_16b	0.024	1.159	2.81	254.60	-0.70	-0.63	-0.07
Comp_16c	-1.092	1.363	3.35	290.24	0.52	0.05	0.47
Comp_16d	1.240	1.539	3.11	264.04	-1.00	-1.24	0.24
Comp_16g	-0.189	1.409	3.51	295.94	-1.40	-0.68	-0.72
Comp_16e	-0.411	1.565	3.36	284.76	0.60	0.17	0.43
Comp_16f	-0.449	1.519	4.36	328.77	0.30	0.11	0.19
Comp_16k	-0.497	1.688	3.36	327.01	-2.18	-2.26	0.08
Comp_7d	-0.054	1.306	3.36	323.924		-2.50	
Comp_9	-2.265	-2.144	3.36	325.078		-2.49	
Comp_15b	-0.866	1.320	2.03	265.248		-2.13	
Comp_15c	-0.992	1.263	2.83	297.630		-1.67	
Comp_15d	1.240	1.539	2.03	294.524		-5.44	
Comp_16h	-0.106	1.719	4.05	343.939		-1.49	
Comp_16i	-0.331	1.524	3.81	320.908		-1.33	

Comp\_7d, Comp\_9, Comp\_15b, Comp\_15c, Comp\_16h, Comp\_16i are not used for regression calculation because of no definite activity data available for those compounds.

cant, low molecular surface is favorable to RT activity. So we can see that the interaction between the drug-receptor is really very complex, the effects of electronic features, steric features and hydrophobic features are all very potential.

But the fitness score of the model comprising the above four important features is low and its multiple linear coefficient is only 0.811, cross-validation co-

efficient only 0.520. We think it is understandable. The results from the statistical analysis only tell us statistical results and do not mean the model using

Table 3

The frequency of the features that appeared in the 73 best models

Features	$\pi_{\text{mol}}$	$\pi_{\text{frg}}$	Surf	Volum	Sfrg	E_for
$n_f$	5	34	28	14	14	2
Features	E_ion	Weight	E_homo	E_lumo	Dip_x	Dip_y
$n_f$	5	3	5	12	0	0
Features	Dip_z	Dip	den_7	den_6	den_5	den_9
$n_f$	29	4	18	1	3	19
Features	char_7	char_6	char_5	char_9	den_pi1	den_pi2
$n_f$	3	0	4	10	28	3

$n_f$  = frequency of the features.

Table 4

List of structural parameters of the most important factors

Comp_num	$\log(1/C)$	$\pi_{\text{frg}}$	Surf	dip_z	den_pi1
Comp_6a	-2.00	2.03	254.034	-0.889	6.262468
Comp_7a	-0.70	3.12	263.812	-0.364	6.262202
Comp_7b	-1.18	3.36	276.482	-0.169	6.261190
Comp_7c	-2.78	3.36	293.399	1.072	6.290723
Comp_7e	-2.70	3.36	292.331	-0.269	6.290116
Comp_8	-2.70	3.36	277.817	2.523	6.256459
Comp_15a	-2.00	1.33	210.733	0.690	6.300765
Comp_16a	-1.00	2.66	236.975	0.200	6.304856
Comp_16b	-0.70	2.81	246.723	0.024	6.304469
Comp_16c	0.52	3.35	270.259	-1.092	6.304509
Comp_16d	-1.00	3.11	254.487	1.240	6.302189
Comp_16e	0.60	3.36	269.057	-0.189	6.303014
Comp_16g	-1.40	3.51	282.771	-0.411	6.304023
Comp_16f	0.30	4.36	308.012	-0.449	6.303283
Comp_16k	-2.18	3.36	304.298	-0.497	6.317863

these features must be very good. The statistical analysis from the average may eliminate interaction between some features from a single model and deduce the randomness for a or several QSAR models. But a model with high fitness score only means that it is a good model or it can predict well, it does not mean that the features comprising this model are the most important features, because the fitness score usually can be enhanced by the interaction between some features. The goal of our QSAR study is not only to find the best predicting model but also to find the most major factors and guide us to find more potential compounds. If we cannot grasp the most ultimate features influencing the biological nativity, it is very difficult to guide us to find more potential compounds. So we think a good model from regression analysis and the most important features from the statistical analysis are equally important. This statistical analysis is helpful for QSAR analysis, especially adaptable to the small data set which have not gotten enough information to verify results.

#### 4. Conclusion

In our work, we use genetic algorithm to study the quantitative structure–activity of PBTP and PBP. From this study, we find genetic algorithm is very powerful in QSAR analysis, it offers a new approach to the problem of building activity models. Replacing standard regression analysis with GAs allows the construction of models competitive with or superior to standard techniques and makes available addi-

tional information not provided by other methods. The derived QSAR models in this study were reasonably satisfying based on statistical significance. Using the best model we predict seven compounds whose RT activities are not known. Moreover, from statistically analyzing the results, we got the most important factors that may influence the HIV-1 RT activity. Because of lacking enough RT activities of some compounds we cannot proceed to the deeper analysis of this group of compounds. Our goal is to get some useful information to guide us for further study. Deeper study will be resumed with the development of laboratory work.

The source codes for the QSAR program used in this study are available from the author upon request.

#### Acknowledgements

The work is supported by the National Foundation of Nature Science.

#### References

- [1] D. Rogers, A.J. Hopfinger, *Am. Chem. Soc.* 34 (1994) 854.
- [2] T.J. Hou, J.M. Wang, X.J. Xu, *Chinese Chem. Lett.*, In press.
- [3] J. Friedman, Multivariate adaptive regression splines, Technical report no. 102, Laboratory for Computational Statistics, Department of Statistics, Stanford University, Stanford, CA, Nov 1988.
- [4] G. Canlpiani, V. Nacci, I. Fiorini, M.P.D. Filippis, A. Garofalo, G. Gleco, E. Novellino, S. Altamura, L.D. Renzo, *J. Med. Chem.* 39 (1996) 2672.
- [5] Crippen, *Chem. Inf. Comput. Sci.* 21 (1987) 21.
- [6] J.M. Wang, T.J. Hou, X.J. Xu, *Chinese Chem. Lett.*, In press.